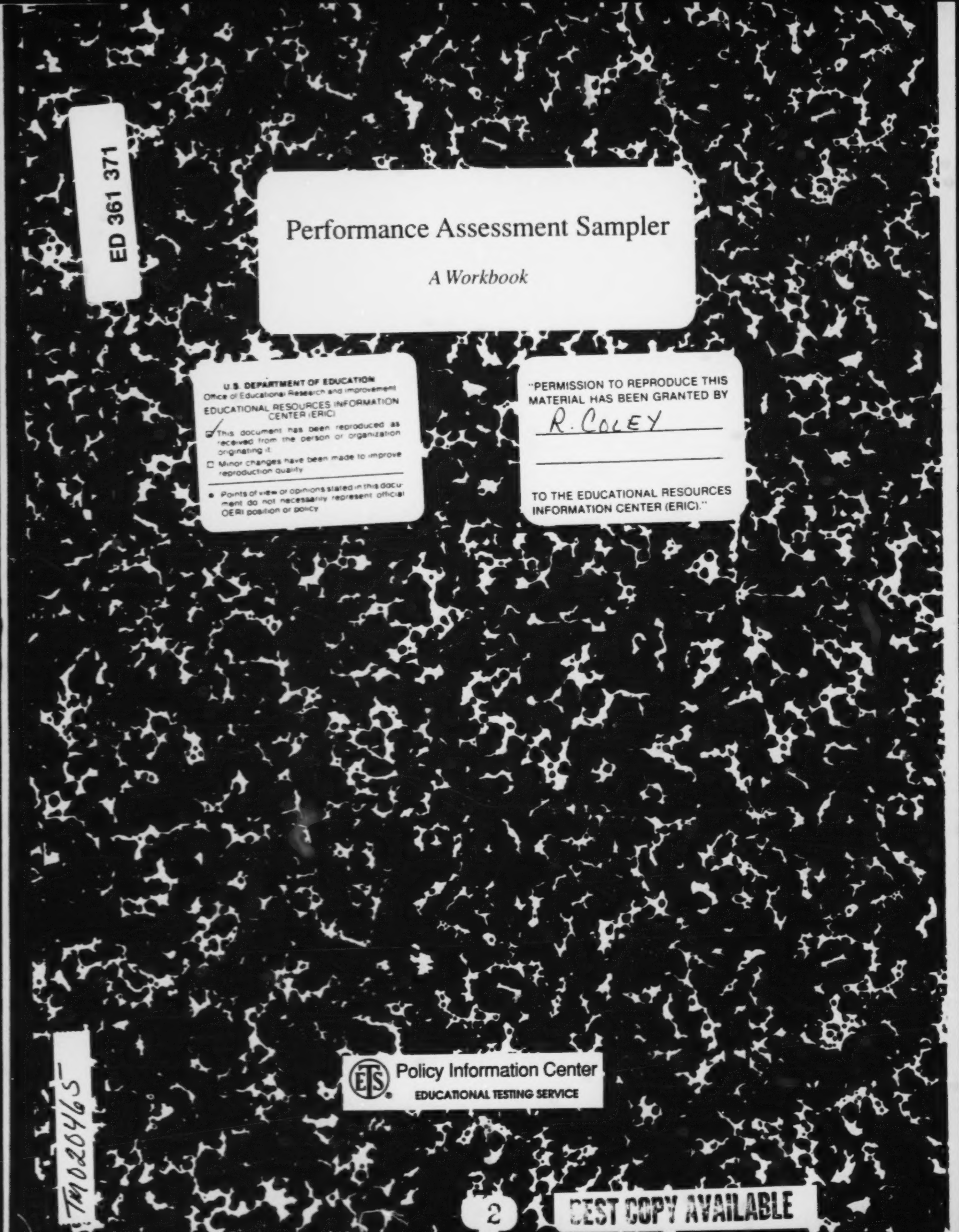
DOCUMENT RESUME

ED 361 371 TM 020 465

TITLE	Performance Assessment Sampler: A Workbook.
INSTITUTION	Educational Testing Service, Princeton, NJ. Policy Information Center.
PUB DATE	93
NOTE	264p.
PUB TYPE	Collected Works - General (020) -- Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC11 Plus Postage.
DESCRIPTORS	Advanced Placement Programs; Cognitive Tests; *Constructed Response; *Educational Assessment; Educational Change; Elementary Secondary Education; High Risk Students; *Portfolios (Background Materials); Problem Solving; Scoring; *Student Evaluation; Test Construction; Test Items; *Thinking Skills; Workbooks
IDENTIFIERS	*Authentic Assessment; Center for Research on Eval Standards Stu Test CA; National Assessment of Educational Progress; *Performance Based Evaluation

ABSTRACT

Performance assessment, constructed-response, and authentic assessment are topics of current interest in educational testing and reform. This workbook presents the following articles on educational assessment to highlight work being done in this area: (1) "Aquarium Problem and Teacher Guidelines: New Standards Project" (University of Pittsburgh, Learning Research and Development Center); (2) "The PACKETS Program" (Educational Testing Service); (3) "Performance Level Guides for Open-Response Common Items, Grade 8" (Kentucky Department of Education); (4) "Advanced Placement Program. 1992 Mathematics: Free-Response Scoring Guide and Sample Student Answers" (Educational Testing Service); (5) "Performance Assessment: Education Research Consumer Guide, Number 2, November 1992" (Office of Educational Research and Improvement); (6) "A Look at a Middle School Portfolio. Arts PROPEL: A Handbook for Visual Arts" (Educational Testing Service); (7) "Multiple Challenges: A Series of Questions Illustrating the 1992 National Assessment of Educational Progress" (Educational Testing Service); (8) "Learning by Doing: A Manual for Teaching and Assessing Higher-Order Thinking in Science and Mathematics" (Educational Testing Service); (9) "NAEP's 1990 Writing Portfolio Study" (Office of Educational Research and Improvement); (10) "Performance Assessment: An International Experiment" (Educational Testing Service); (11) "Measuring What's Worth Learning and Mystery Graphs. Measuring Up: Prototypes for Mathematics Achievement" (National Academy Press); (12) "Piloting Pacesetter: Helping At-Risk Students Meet High Standards" (Thomas W. Payzant and Dennie Palmer Wolf); (13) "Ensuring Reliable Scoring" (Joan L. Herman, Pamela R. Aschbacher, and Lynn Winters); (14) "CRESST Performance Assessment Models: Assessing Content Area Explanations" (Center for Research on Evaluation, Standards, and Student Testing); (15) "The CRESST Line: Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing"; and (16) "Construction versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment" (Randy Elliot Bennett, Ed., and William C. Ward, Ed.). (SLD)



ED 361 371

Performance Assessment Sampler

A Workbook

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. COLEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ETS Policy Information Center
EDUCATIONAL TESTING SERVICE

TM020465

2

BEST COPY AVAILABLE

NAME ______________ ADDRESS ______________

SCHOOL ______________ CLASS ______________

		PERIOD 1 PERIOD 2 PERIOD 3 PERIOD 4 PERIOD 5 PERIOD 6 PERIOD 7 PERIOD 8	
		Introduction and Acknowledgments	**i**
MONDAY	SUBJECT	**Aquarium Problem, New Standards Project**	**1**
	ROOM	**PACKETS™**	**13**
	INSTRUCTOR	**Kentucky Open Response Items**	**31**
TUESDAY	SUBJECT	**Advanced Placement Calculus**	**43**
	ROOM	**OERI Consumer Guide**	**53**
	INSTRUCTOR	**Arts PROPEL**	**59**
WEDNESDAY	SUBJECT	***Multiple Challenges* (NAEP)**	**71**
	ROOM	***Learning by Doing* (NAEP)**	**85**
	INSTRUCTOR	***NAEP's 1990 Writing Portfolio Study***	**119**
THURSDAY	SUBJECT	**International Science Tasks (IAEP)**	**145**
	ROOM	**From *Measuring Up* (Mathematical Sciences Education Board)**	**167**
	INSTRUCTOR	**"Piloting Pacesetter"**	**189**
FRIDAY	SUBJECT	**From *A Practical Guide to Alternative Assessment* (ASCD)**	**195**
	ROOM	**From *CRESST Performance Assessment Models***	**213**
	INSTRUCTOR	***The CRESST Line*, Portfolio Issue**	**229**
		From *Construction Versus Choice in Cognitive Measurement*	**243**

Copyright © 1993 by Educational Testing Service. All rights reserved.

Introducing the Performance Assessment Sampler

Performance assessment, constructed-response, and authentic assessment are terms sweeping through educational testing and reform. Much is in development, in prototype, and in early use. Every day an educator or testing official somewhere is likely being told to get on top of what is being done in this area and get started in their state, district, or school. With the recent spurt in development, that isn't easy to do quickly.

This "sampler" is designed for the person who needs to get a handle on these new assessment efforts. It follows the pattern of the ETS Policy Information Center's previous "workbook" on national educational standards*, reproducing excerpts that give at least an acquaintance with a project, and information on where to go to learn more. This is by no means an exhaustive inventory of efforts at alternative assessment going on in the United States, or at Educational Testing Service. Rather, it is a sampler that attempts to represent a broad range of efforts in this area.

Paul E. Barton
Director

Richard J. Coley
Senior Research Associate

Acknowledgments

We are indebted to all of the individuals and organizations who gave us permission to reproduce their materials. This permission, when required, is specified on the page introducing the particular project or material.

Carla Cooper provided desktop publishing services and preparation for printing.

***National Standards for Education: What They Might Look Like. A Workbook*. Princeton, NJ: Policy Information Center, Educational Testing Service, 1992.

Aquarium Problem and Teacher Guidelines

New Standards Project

The New Standards Project is a joint program of the National Center on Education and the Economy, which is based in Rochester, NY, and the Learning Research and Development Center at the University of Pittsburgh. The project has attracted the participation of 17 states and six large school districts who already were far along in designing and administering a new generation of assessments based on performance rather than multiple-choice tests.

The system created by the Project will set a high standard of performance for all students. The assessments will emphasize the ability to think well, to demonstrate a real understanding of subjects studied and to apply what one knows to the kind of complex problems encountered in life. The Project will employ portfolios, exhibitions, projects, and timed performance examinations, all based on the use of real-life tasks that students are asked to do alone and in groups.

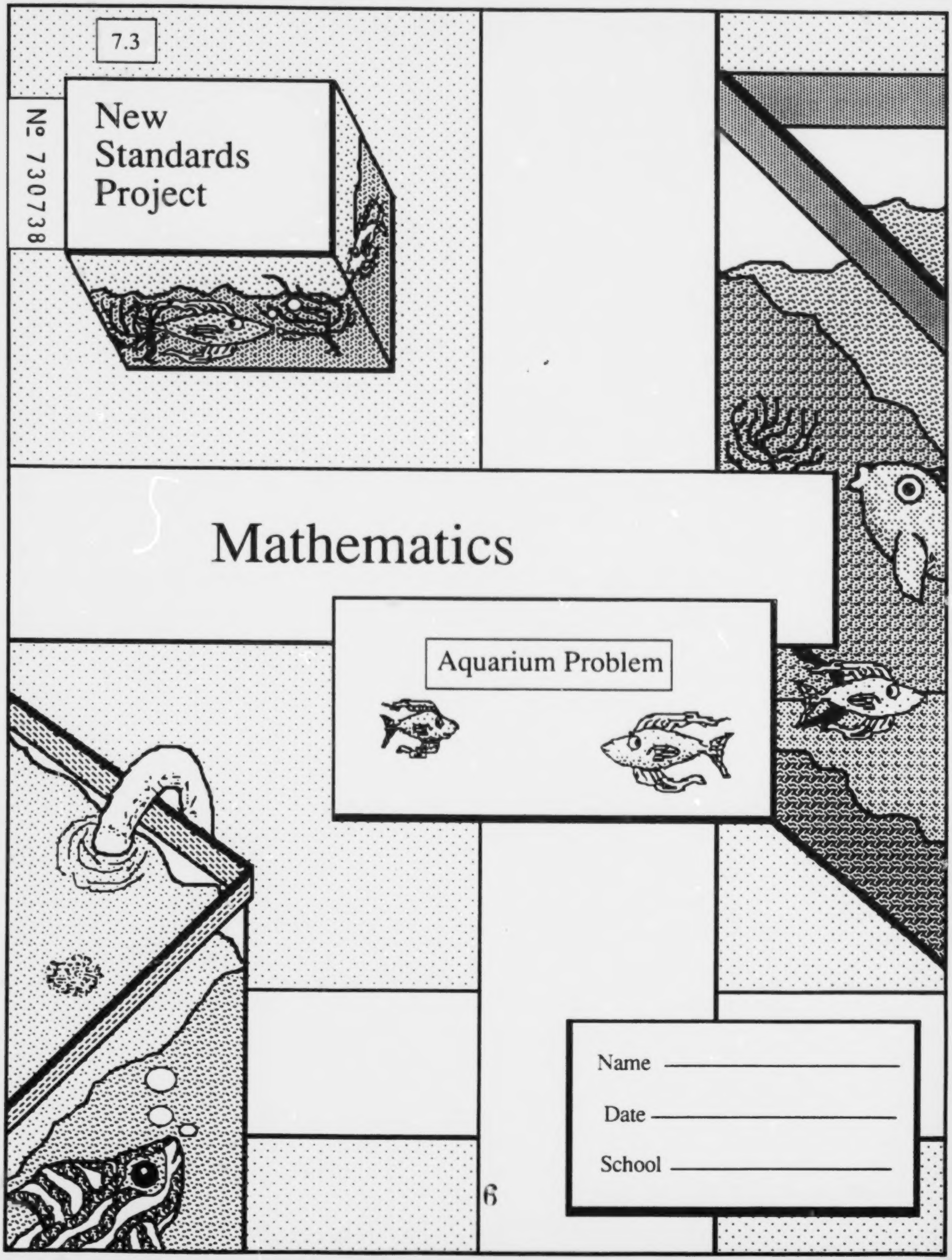
In establishing content standards, the Project is drawing on the work of national bodies such as the National Council of Teachers of Mathematics and on curriculum frameworks and goals developed by the states. It will also work to establish international benchmark standards for student performance. Work has begun on the tasks that will constitute the core of the examinations and the first exams will be available in 1993-94.

For more information on the New Standards Project, write to:

Learning Research and Development Center
University of Pittsburgh
3939 O'Hara St., Room 408
Pittsburgh, PA 15260

or call 412-624-8319

The "Aquarium Problem" is reproduced with the permission of the New Standards Project. The fish illustrations are reproduced with the permission of T.F.H. Publications, Inc., Neptune, New Jersey.



7.3

Nº 730738

New Standards Project

Mathematics

Aquarium Problem

Name ______________________

Date ______________________

School ______________________

NEW STANDARDS PROJECT

a joint program of the
Learning Research and Development Center at the University of Pittsburgh
and the
National Center on Education and the Economy

May 15, 1992

Dear Student:

Today you will be part of an exciting plan called the New Standards Project. We are looking at new ways of teaching, learning and testing. Our plan is to create interesting learning activities for students. We hope that these activities will give you a chance to show what you know and what you can do in math.

All across the country, fourth graders from many communities are helping the New Standards Project by working on these learning activities. By showing and explaining your best thinking, you will help us improve the activities before trying them with other students.

We thank you for your help and for being such an important part of the New Standards Project.

Sincerely,

Philip Daro
Director for Mathematics

BEST COPY AVAILABLE

•
New Standards Project
Learning Research and Development Center University of Pittsburgh
Room 408 3939 O'Hara Street Pittsburgh, PA 15260 Tel: 412-624-8319 Fax: 412-624-9149
•
New Standards Project
National Center on Education and the Economy
39 State Street Suite 500 Rochester, NY 14614 Tel: 716-546-7620 Fax: 716-546-3145
•

THE AQUARIUM

Imagine that your school principal asks you to do a special job and gives you these written directions:

Your class will be getting a 30 gallon aquarium. The class will have $25.00 to spend on fish. You will plan which fish to buy. Use the Choosing Fish for Your Aquarium brochure to help you choose the fish. The brochure tells you things you must know about the size of the fish, how much they cost and their special needs.

Choose as many different kinds of fish as you can. Then write a letter to me explaining which fish you choose. In your letter,

1. tell me how many of each kind of fish to buy
2. give the reasons you chose those fish
3. show that you are not overspending and that the fish will not be too crowded in the aquarium.

Choosing Fish for Your Aquarium

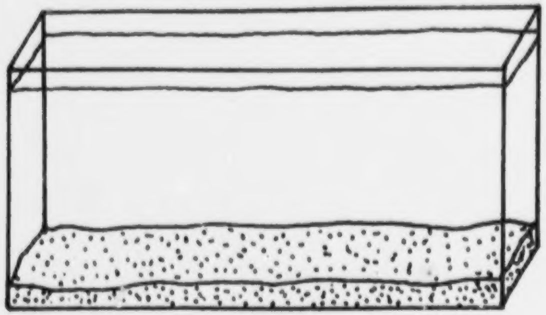
Planning Ahead

Use the information in this brochure to help you choose fish that will be happy and healthy in your aquarium. To choose your fish, you must know about the **size** of the fish, their **cost**, and their **special needs**.

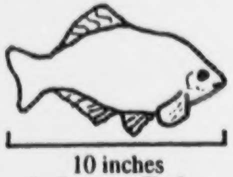
Size of Fish

To be healthy, fish need enough room to swim and move around. A good rule is to have one inch of fish for each gallon of water in your aquarium. This means that in a ten gallon aquarium, the lengths of all your fish added up can be ten inches at the most.

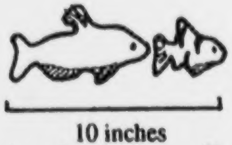
EXAMPLE:
With a ten gallon aquarium,



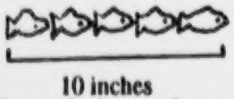
here are a few of your choices:



one ten-inch long fish, **or**



a seven-inch long fish and a three-inch long fish **or**



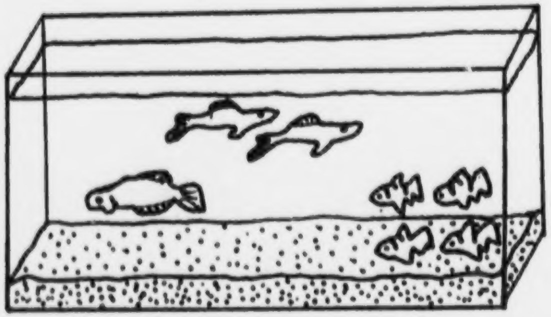
five fish if each is only two inches long.

Cost of the Fish

Some fish cost as little as one dollar, others cost much more. The prices of each kind of fish are listed in the chart.

Special Needs

Use the chart to learn about the special needs of each kind of fish. Some fish need to live together in **schools** --a group of four or more of the same kind of fish -- while other live in pairs or alone. A few kinds of fish have other special needs, which are listed in the chart.



Student Reflections, Ideas

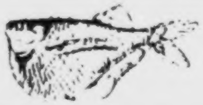
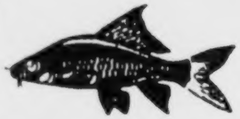
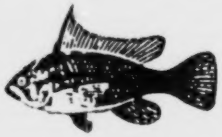
You can help us make these learning activities even better. Think about each of the following questions, and write to us what you honestly think. Be as clear as you can (you might want to give us examples of what you mean).

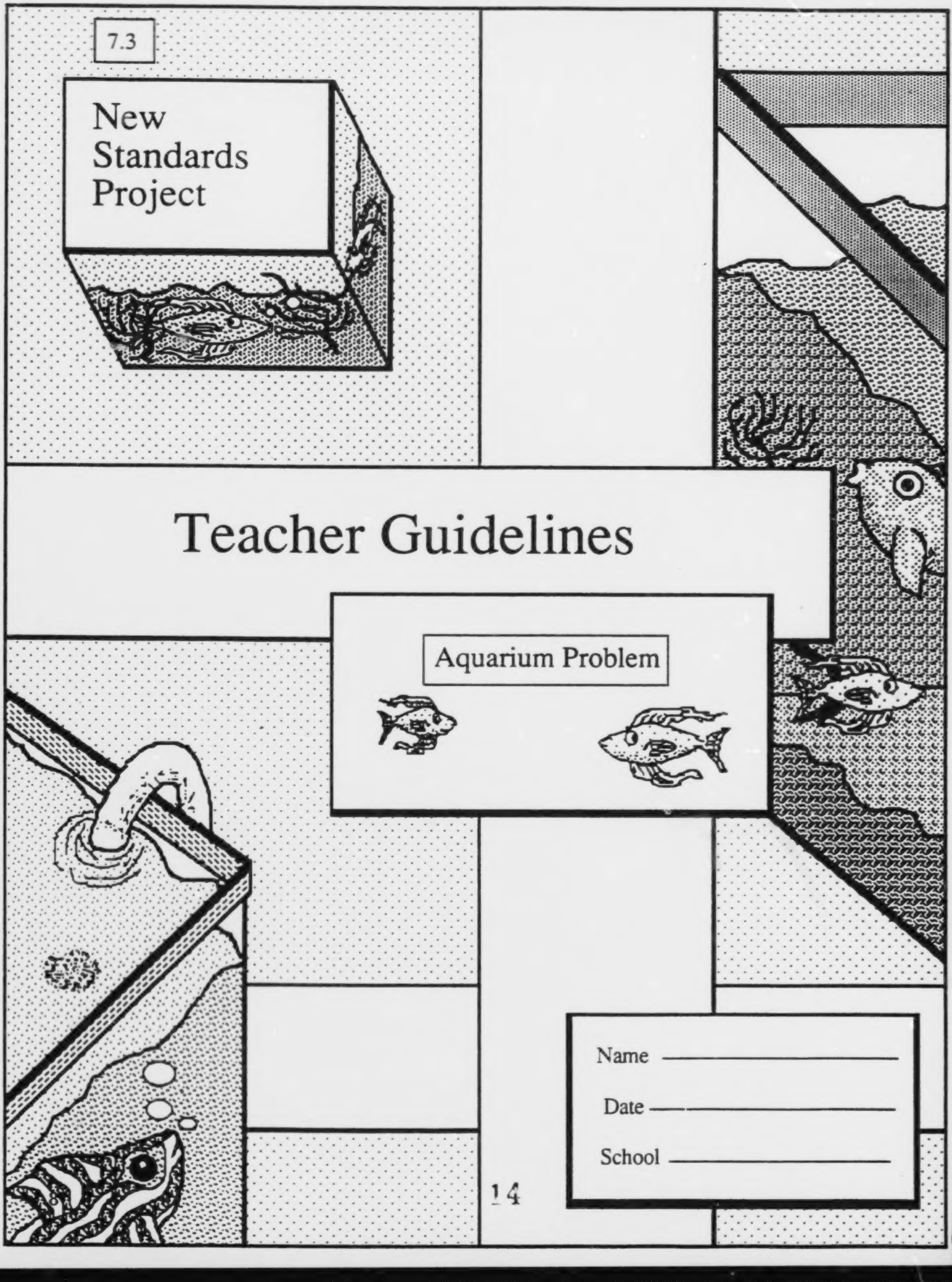
What did you enjoy about the task?

What did you not like about the task?

How is this task like other activities you do in your class? How is it different?

Chart for Freshwater Fish

Picture	Name	Cost	Length in Inches	Color	Special Needs, Facts
	Zebra Danio	$1	$1\frac{1}{2}$ inches	blue with gold lines	Lives in schools; gets along with other kinds of fish.
	Marbled Hatchetfish	$1	2 inches	yellow	Lives in schools; can leap 3-5 yards.
	Guppy	2 for $3	2 inches	red, blue and green	Lives in schools.
	Red-tailed Black Shark	$5	$4\frac{1}{2}$ inches	black with red tail	Fights with other sharks, but gets along with other kinds of fish.
	Cardinal Tetra	$5	$1\frac{1}{2}$ inches	red and green	Lives in schools.
	Blind Cave Fish	$2	3 inches	silvery rose	Lives in schools; uses its sense of smell and vibration to find food.
	Ramirez' Dwarf Cichlid	$5	2 inches	rainbow	Lives in pairs; rarely lives longer than $2\frac{1}{2}$ years; gets along with other fish.
	Velvet Cichlid	$5	$12\frac{1}{2}$ inches	olive with stripes	Can be trained to take food from the hand and can be petted. Must be kept only with other cichlids



7.3

New
Standards
Project

Teacher Guidelines

Aquarium Problem

Name ____________________

Date ____________________

School ____________________

THE AQUARIUM
Guidelines for the Teacher

Purpose

In **The Aquarium** task, students use their knowledge of mathematics to solve a real world problem. Students use logical and numerical reasoning about money, measurement and realistic conditions to decide how best to stock an aquarium within the constraints of the situation.

This task calls for logical and numerical reasoning and justification of that reasoning. These are contained in Standard 1 (Mathematics as Problem Solving), Standard 3 (Mathematics as Reasoning), and Standard 4 (Mathematical Connections) of the NCTM Curriculum Standards for Grades K-4. Students apply understanding of measurement as described in Standard 10 (Measurement).

This task is designed to assess students' mathematical thinking and their use of information contained in a brochure which features a chart, not their ability to read the brochure and chart independently.

Materials and Resources

- Students should have easy access to calculators
- Paper, pencils
- Extra copies of Choosing Fish for Your Aquarium brochure and Chart for Freshwater Fish (the last page of the task booklet)
- Task booklet

Time Required

The Introductory Activities take about one 40-minute class period; anticipate two class periods for the assessment itself. Allow enough time for the task so that you and the students feel that they have been able to do their best work.

Ideas for post-assessment activities are included at the end of these guidelines. You may wish to use these to extend the exploration of the mathematical ideas contained in the assessment task.

Introductory Activity

Introductory activities should interest students in the task context (fish and aquariums) and ensure that all students have sufficient knowledge and curiosity about the context to enable them to work on the task. You may want to start by helping the whole group generate and discuss some considerations in selecting fish for an aquarium. Guide students to include these criteria: price of fish, choosing types of fish that can live together, allowing sufficient room and oxygen for all the fish.

After the discussion, help students read the Choosing Fish for Your Aquarium brochure. Help them understand the chart and help them to practice with it. Have students notice and tell all that they can about a particular fish listed on the chart. Have students find fish that must live in groups of four or more ("schools"). Other possible questions: "My friend has a fish with red on it, what could it be?" "How much will it cost to buy four guppies?" "What is the shortest fish? The longest?" "Is there a yellow fish that is longer than 2 inches?" "Which fish is the most expensive."

To build students' interest and readiness, a Think-Pair-Share activity can be helpful:

- Write the words "fish" and "aquarium" on the chalkboard. Ask students to think about these words, then to write down everything that they know about "fish" and "aquarium". Have students write down complete ideas, not just single words.
- Allow students to form pairs and share their lists.
- Now ask students to share as an entire group. Record their ideas and keep the group's list visible while students work on the assessment task.

Assessment Task

Read the Letter to the Students at the beginning of the task. This should help communicate that a good effort is expected of all students.

Introduce the **Aquarium** task. You may want to refer back to your introductory activities. Read the task aloud with the students, or read it to them. Explain any words you think may be unclear. Make sure that students understand the considerations for choosing fish and that they know what to include in their letter to the principal.

Each student should work independently to create a workable solution, and write a letter explaining his or her choices. Your interaction with the students should be limited to making them comfortable with the assignment and to normal classroom management.

When the students have finished the assessment task, ask them to respond to the **Student Reflections, Ideas** questions on the last page of the task booklet.

Post-Assessment Suggestions

The purpose of the post-assessment activity is to provide students with an opportunity to review how they solved the problem and to learn from their work on this task.

Suggested Activities:

- A "pair-share" procedure, similar to the one in the introductory activity, is one technique for doing such a reflective review. Students share solutions with their classmates, thinking about the crucial elements of the task, revise or at least revisit their work, and finally, reflect upon what they think they learned as a result of participating in this activity.

- In this activity, students share solutions with their classmates, thinking about the crucial elements of the task, revise or at least revisit their work, and finally, reflect upon what they think they learned as a result of participating in this activity.

Begin by talking with students about how we often become better problem solvers by reviewing how we and others solved a particular problem. Mention that everyone is to be commended for the effort they put forth in working on this task. In order for us to improve our own problem solving skills, we are going to share our work.

The activity could continue something like this:

> "First, we'll share with partners. Each of you will exchange your plan with a partner. As you read your partner's solution to the problem, note at least two things that you think showed good thinking. Also, write at least one or two questions which you might ask to better understand what your partner did to solve the problem or which might help your partner improve his solution to the problem Share your responses with your partner."

To help students review each other's work, remind them of the *critical task parameters*--30 gallon tank and $25.00 limit--and the important information about the fish to be chosen-size, cost, and special qualities.

Have students rotate partners within their groups. Students will need to keep track of the good points and questions/suggestions for each partner.

After students have had an opportunity to share with their partners, reconvene the entire class. Pose questions such as:

> What surprising or interesting things did you learn?
> What else would you like to share with the class?
> What would you do differently if you were to revise your solution for the task?

You might then ask students go back and actually revise their solutions. Finally, you might ask students what they learned from working on the Aquarium task. Students should record their responses.

The PACKETS™ Program

PACKETS™ is a major new program of Educational Testing Service to develop performance-based activities that teachers can use as part of classroom instruction and as the foundation for documenting the learning process.

The program contains a series of high quality, nationally field-tested performance assessment activities or tasks. These materials are packaged by specific subjects and grade levels. PACKETS™ materials include activities that:

- Do not presume one correct way of thinking about the problem or just one right answer.
- Require students to utilize a broad spectrum of knowledge, reasoning, problem-solving, and communication.
- Require students to work in groups in a cooperative learning environment.

Although the PACKETS™ program will cover the K-12 spectrum across several content areas, the first set of materials is in middle school mathematics. The Middle School Math PACKETS™ program is currently in use in a limited number of field-test classrooms, and will be available nationally for the 1994-95 school year.

The materials provided here include some overall information on the program, along with examples of activity, feedback, and assessment materials.

For more information, contact:

Nancy Katims
Mail Stop 37-B
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

These materials are copyrighted by Educational Testing Service and are reproduced here with permission.

The chart, "Race of the Sexes: What Lies Ahead," presented in the "PACKETS Times," is copyrighted by the New York Times. Reprinted by permission.

The PACKETS™ Program

Educational reform in the 1990's is calling for change – change in classroom instructional practices and change in testing and assessment. In order for true educational improvement to occur, assessment and instruction must be linked. A strong focus on higher-level thinking and on interdisciplinary activities must exist.

One of the Educational Testing Service's major undertakings is the development of performance-based activities that teachers can use as part of classroom instruction and as the foundation for documenting the learning process. This concept is reflected in the PACKETS™ Program.

What Is the PACKETS™ Program?

The PACKETS™ program contains a series of high quality, nationally field-tested performance assessment activities or tasks. These materials are packaged by specific subjects and grade levels. PACKETS™ materials include activities that:

- ▲ do not presume one correct way of thinking about the problem or just one right answer,
- ▲ require students to utilize a broad spectrum of knowledge, reasoning, problem-solving, communication, and "big ideas," where skills and small or isolated bits of knowledge are placed in a larger picture, and
- ▲ require students to work in groups in a cooperative learning environment to create a product which meets the requirements of the activity.

In addition to the activities or tasks, the PACKETS™ program contains information on how to interpret a student's response as well as how to foster self-evaluation on the part of the student.

PACKETS™ materials furnish teachers with the tools they need to implement performance assessment, while stimulating students' thinking and gettting students excited about learning. Students learn how to evaluate their own responses and to revise and improve their own work. Assessment, therefore, becomes an important part of the learning process.

Math PACKETS™ Program

Although the PACKETS™ program will cover the K-12 spectrum across several content areas, the first set of PACKETS™ materials is in middle school (grades 6-8) mathematics. The framework underlying the materials is built upon many years of research conducted by Richard Lesh and others in the mathematics education field.

In the math model, each "big ideas" project is based on a real-life data-rich newspaper article carefully written for middle school students. Accompanying each project is a sequence of supporting activities designed to help students internalize and apply the mathematical ideas generated in the project activity. Included within each sequence of tasks are problems suitable for in-class group and individual work as well as for homework.

The math PACKETS™ materials are strongly interdisciplinary, with activities that encourage reading, writing, oral communication, and mathematical skills, reasoning and problem-solving.

The Middle School Math PACKETS™ program is currently in use in a limited number of field-test classrooms, and will be available nationally for the 1994-95 school year.

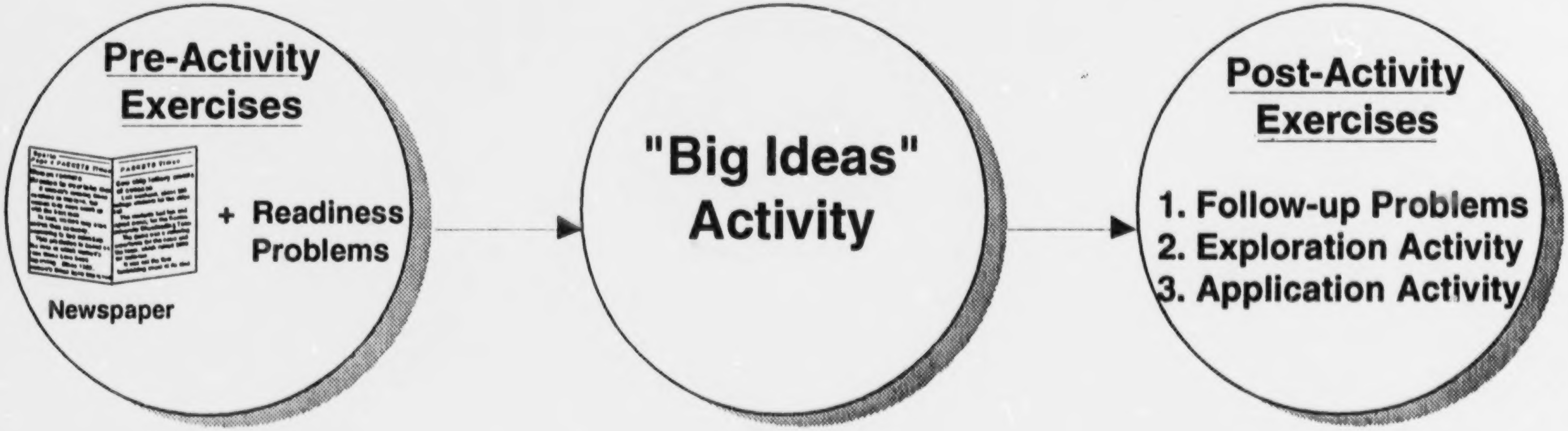
What Are The Key Features?

An individual subject matter and grade level set of PACKETS™ materials includes:

- ▲ **Teachers' Guide**
  A wealth of information on a variety of topics, including the value and utility of performance assessment, the role of performance assessment in classroom teaching and evaluation, how to integrate these activities into one's teaching, and how to use students' work from PACKETS™ activities to build a portfolio.
- ▲ **Classroom Tested Performance Assessment Activities**
  Project-sized activities in which students bring together skills, reasoning, and problem-solving strategies to create a solution to a problem that has real-world application. Written and oral expression are components across all subjects. Since the activities are field tested by a sample of teachers and students across the country, the activities represent materials that work.
- ▲ **Sample Student Responses**
  Examples of student products, from across the country, which represent a variety of approaches to a given "big ideas" project. These sample responses are selected to assist the teachers and students in understanding and applying the principles of the assessment guide to the products developed in their own classrooms.
- ▲ **Assessment Guide**
  A set of tools designed to help teachers and students capture the richness and multi-dimensional nature of the student products at both a descriptive and an evaluative level. Reflecting the high quality assessment expertise of ETS staff coupled with input from teachers and educational leaders nationally, this component provides guidelines and models for the development of feedback mechanisms that can inform multiple audiences for instructional and assessment purposes.

Copyright © 1993 by Educational Testing Service. All rights reserved.

MODEL OF A PACKETS™ ACTIVITY SEQUENCE FOR MIDDLE SCHOOL MATHEMATICS



Pre-Activity Exercises

Newspaper + Readiness Problems

"Big Ideas" Activity

Post-Activity Exercises

1. Follow-up Problems
2. Exploration Activity
3. Application Activity

Interpretation Framework

Sample Student Products

Description of Mathematical Approaches

Instructional Feedback Options

Assessment Guidelines

Copyright © 1993 by Educational Testing Service. All rights reserved.

PACKETS Times

*A resource for
developing and demonstrating
mathematical power
through performance activities*

Copyright © 1993 by Educational Testing Service. All Rights and Occasional Wrongs Preserved.

Vol. 2, No. 1. Spring 1993.

Women runners threaten to overtake men

If women's running times continue to improve, top women may soon catch up with the best men.

In fact, women may even outrun men someday, according to two scientists.

This prediction is based on the rate at which women's race times have been improving. Since 1920, women's times have improved much more than men's.

Resarchers say that the best female runners should run marathons as quickly as men by 1998. Women should catch up with men in shorter track events by the middle of the next century.

These predictions are based on a comparison of trends in men's and women's world records over the past 70 years. Based on these patterns, projections are made into the future.

The results were published by Dr. Brian J. Whipp and Dr. Susan A. Ward in the British journal, *Nature*. The two scientists teach at the University of California at Los Angeles.

Dr. Peter Snell, an exercise physiologist at the University of Texas, does not accept the results. "I'd agree that there's a way to go yet in women's performance, but if they're suggesting that women will approach men, that's ludicrous."

But the two researchers said the women's trend has been too consistent to ignore.

Whipp said that before looking at the data, he did not think women would ever catch men. But now, he thinks, "Men and women might be running equivalent speeds in the next century."

He added, "This is not me talking. It's the data."

In 1954, when Roger Bannister became the first man to run a four-minute mile, Diane Leather became the first woman to run a five-minute mile. If they had been in the same race, she would have finished 320 meters behind Bannister.

Today, the top female runner would finish only 180 meters behind the fastest man, according to Whipp.

In the marathon, women's times have improved about 61 percent since 1955. Men's performance has improved only 18 percent.

Women have come a long way, but there is still a long way to go to catch up. The fastest female runners today would not even qualify for the men's track events in the Olympics.

In the marathon, the men's world record is 2:06.50, while the women's is 2:21.06. By marathon standards, this is a huge difference.

Many people doubt that women will ever catch up.

"Women will never, ever catch up to the men," said Frank Lebow, president of the New York Road Runners Club. "Maybe on paper this looks good, but I've been to 100 marathons around the world, and I've seen all the women runners. Women will never pass men. Never, never."

Joan Benoit Samuelson, the 1984 Olympic marathon champion and record-holder among American women, said that women might get closer to men's times, but would never beat them.

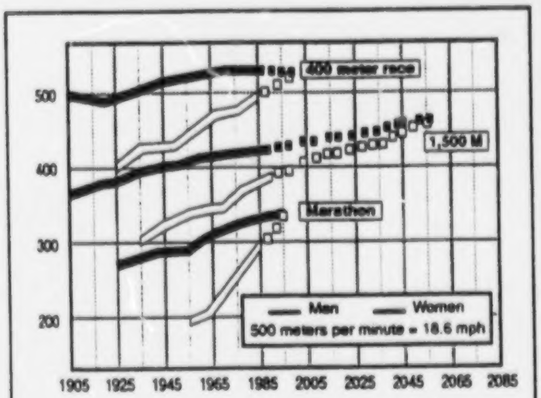
"Men have had a lot more time to evolve in the sport, and since they've got that jump start, they'll be hard to beat now," she said. "You also have the male ego to consider, and that's going to keep men going."

According to Snell, who won three Olympic gold medals in the 1960s for running, men also have physical advantages. Men have larger muscles, stronger bones, and a smaller percentage of body fat. Men also have more red blood cells, so they can get more oxygen to their muscles.

Snell thinks women's improvements are due to social changes. "Finally, women are starting to get out and do the same things as men."

Patti Sue Plumer, who in 1990 was ranked No. 1 in the 3000-meter and 5000-meter events, does not think that physical advantages are that important. "As an athlete, I've learned that the mind plays a much stronger role than anything physiological," she said.

Perhaps the debate will be settled only by time. For female runners, the race has only just begun.



Race of the Sexes: *What Lies Ahead?*

Average running velocity for record-holders of both sexes in major events in meters per minute. The historical trend is extrapolated to the point where women would catch up to men.

This special issue of the PACKETS Times was published as part of an ETS project on newspaper-based performance activities for mathematical instruction and assessment.

EDUCATIONAL TESTING SERVICE, ETS and the ETS logo are registered trademarks of Educational Testing Service.
PACKETS™ is a trademark of Educational Testing Service. Copyright © 1993 by Educational Testing Service.
All rights reserved.

"Big Ideas" Activity

PACKETS™ Project
The Fast Track

Do you think that women may soon outrun men? How fast do you think women and men will run in 100 years? In 200 years?

The editor of the school newspaper has decided to write an article titled "The Fast Track." The article will include predictions and comparisons of the speeds for women and men in the 200-meter run of future Olympic Games. The editor has asked your class to predict what the speeds might be for the next 50 Olympic Games (the next 200 years).

Write up your predictions and conclusions for the editor. The editor will need to explain and justify the predictions in the article. Therefore, include any graphs, charts, or other materials that would help the editor understand the reasoning for your predictions.

Gold Medalists in the Women's 200-Meter Event

Year	Name, Country	Time in seconds	Speed in mph
1988	Florence Griffith-Joyner, United States	21.34	20.9
1984	Valerie Brisco-Hooks, United States	21.81	20.5
1980	Barbel Wockel, E. Germany	22.03	20.3
1976	Barbel Eckert, E. Germany	22.37	20.0
1972	Renate Stecher, E. Germany	22.40	19.9
1968	Irena Szewinska, Poland	22.5	19.8
1964	Edith McGuire, United States	23.0	19.4
1960	Wilma Rudolph, United States	24.0	18.6
1956	Betty Cuthbert, Australia	23.4	19.1
1952	Marjorie Jackson, Australia	23.7	18.8
1948	Francina Blankers-Koen, Netherlands	24.4	18.3

Copyright © 1993 by Educational Testing Service. All rights reserved.
These materials are provided for the sole purpose of classroom tryouts. Any use, reproduction, or distribution for any other purpose requires the prior written consent of the ETS Director of School Services.

PACKETS™ Project
The Fast Track

Gold Medalists in the Men's 200-Meter Event

Year	Name, Country	Time in seconds	Speed in mph
1988	Joe DeLoach, United States	19.75	22.6
1984	Carl Lewis, United States	19.80	22.5
1980	Pietro Mennea, Italy	20.19	22.1
1976	Donald Quarrie, Jamaica	20.23	22.1
1972	Valeri Borzov, USSR	20.00	22.3
1968	Tommie Smith, United States	19.83	22.5
1964	Henry Carr, United States	20.3	22.0
1960	Livio Berruti, Italy	20.5	21.8
1956	Bobby Marrow, United States	20.6	21.7
1952	Andrew Stanfield, United States	20.7	21.6
1948	Mel Patton, United States	21.1	21.2
1936	Jessie Owens, United States	20.7	21.6
1932	Eddie Tolan, United States	21.2	21.1
1928	Percy Williams, Canada	21.8	20.5
1924	Jackson Scholz, United States	21.6	20.7
1920	Allan Woodring, United States	22.0	20.3
1912	Ralph Craig, United States	21.7	20.6
1908	Robert Kerr, Canada	22.6	19.8
1904	Archie Hahn, United States	21.6	20.7
1900	Walter Tewksbury, United States	22.2	20.1

Copyright © 1993 by Educational Testing Service. All rights reserved.
These materials are provided for the sole purpose of classroom tryouts. Any use, reproduction, or distribution for any other purpose requires the prior written consent of the ETS Director of School Services.

Examples of Pre-Activity Readiness Problems for PACKETS™ Middle School Math

1. For each year during the past six years, estimate how fast you have run. Sketch a graph of your running speeds across those years.
2. In a sentence or two, write a definition of a trend. Provide an example. Then draw a graph that shows this trend.
3. A friend claims that 550 meters per minute is the same as 18.6 mph. Write a sentence or two that either justifies or disputes your friend's claim.
4. Approximately how fast did women run the marathon in 1955?
5. If you rode a bike 100 yards in 20 seconds, what is your speed in miles per hour? What is your speed in meters per minute?
6. If you rode a bicycle faster than any man has run a marathon, but slower than the fastest woman has ever run 1500 meters, how fast might you be riding?

Examples of Post-Activity Exercises for PACKETS™ Middle School Math

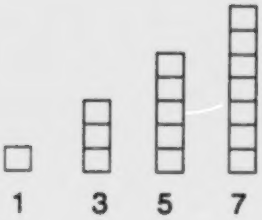
I. Follow-Up Problems

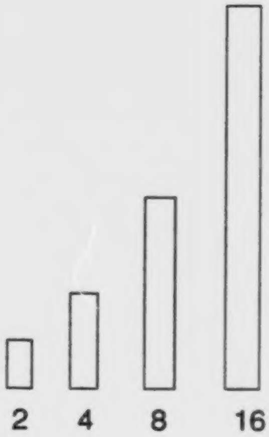
1. a. Using the information that is given in the graph in the newspaper, predict the running speed of the 1992 world record holder for each of the three events.
   b. Compare your predictions with the actual 1992 world records.
   c. In a sentence or two, describe any significant differences between your predictions and the actual records.
2. If you ran the 200 meter event in 21 seconds, what would your average speed be?
3. a. What is the average difference in running speeds of the gold medalists in the men's and women's 200 meter Olympic events for the Olympics between 1948 and 1988?
   b. Graph the average difference.
4. Describe in your own way the data given for the gold medalists in the men's 200 meter event without giving any of the actual numbers in the description.
5. In doing this activity, what tools or resources would you have liked that might have been helpful? Describe in a few sentences how these tools might have changed your predictions.
6. By how many percentage points did the running speeds of the gold medalists in the women's 200 meter Olympic event increase between 1948 and 1988?

Copyright © 1993 by Educational Testing Service. All rights reserved.

II. Exploration Activity

The exploration activity is an activity in which the students explore the pure mathematics of the activity in a concrete or graphic representation. For "Fast Track," the students might look for patterns in sequences of the following types:

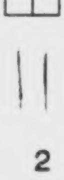


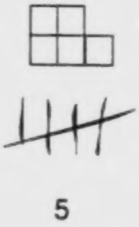


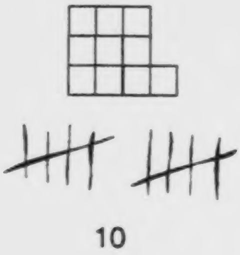
1	②	3	4	5	⑥	7	8	9
10	11	⑫	13	14	15	⑯	17	18
19	⑳	21	22	23	㉔	25	26	27
28	29	㉚	31	32	33	㉞	35	36
37	㊳	39	40	41	42	43	44	45
46	47	48	49	50	51	52	....	

2, 6, 12, 16, 20 ,24, 30, 34, 38,...

1,②3, 4,⑤6, 7, 8, 9, ⑩ 11, 12, 13, ...







III. Application Activity

The application activity encourages the students to extend the ideas developed in the "big ideas" activity. For example, the application activity for "Fast Track" might ask the students to make projections in a new content area such as world population growth.

Copyright © 1993 by Educational Testing Service. All rights reserved.

Description of Mathematical Approaches for the "Fast Track" Activity

Units of Analysis (Simple vs Composite)

What are the units people think about when working on this problem? Sometimes people use small simple units such as one year, one Olympics, individual running speeds or individual running times. Other times they use larger, composite units, such as blocks of data, patterns or trends.

[illegible] meter [illegible]
women gain .197 seconds every year.
[illegible] women will surpass men

Time In Sec	Speed in mph
19.75 ± .05	22.61 ± .1

Differences vs Ratios

How do people think about change? Sometimes people think of change in terms of differences (absolute change). Other times they think more in terms of percentages or ratios (relative change). Change can be relative to time intervals (e.g., years) or change can be relative to running speeds or running times. Complex ratios can be relative to both time intervals and running speeds/times.

[illegible] from 1942 – 1988
men improved by 1.35 yard
women [illegible] by

Men on average have increased at .8 mph per year [illegible]

$\frac{.1}{3}$ = increase by 14.2%
increase by 6.6%

EDUCATIONAL TESTING SERVICE, ETS and the ETS logo are registered trademarks of Educational Testing Service.
PACKETS™ is a trademark of Educational Testing Service. Copyright © 1993 by Educational Testing Service.
All rights reserved.

Little-Picture vs Big-Picture

How do people think about making predictions based on past performances? Sometimes people think about the problem by projecting from little-picture information what a next one will be, then a next one, a next one, and so on. Other times they extrapolate from big-picture information what some future situation will be and then attempt to fill in the holes.

every year women run the 200 meter .326 faster.

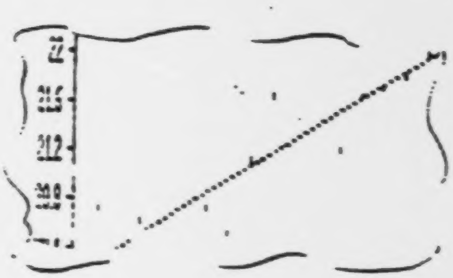
in 40 years women gained 1.71 seconds, so in 5 x 40 years (200 years) men will be behind

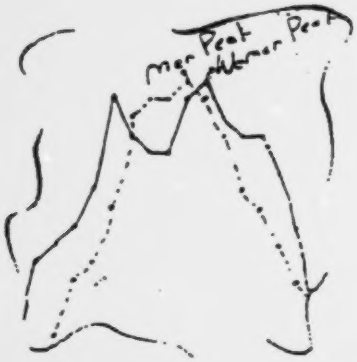
Linear vs Non-Linear

How do people think about trends? Sometimes people think about the problem in a linear fashion. They see a constant rate of increase and project it in a steady-state fashion. Other times people think about the problem in non-linear ways. They think in terms of a limiting factor or a dynamic rate of change. This may be expressed as leveling off, maxing out or peaking.

See the leveling off of times and rates → such a trend which men have followed..

Women will eccellerate to a degree but w probably still be in the 19 to 23 second range.

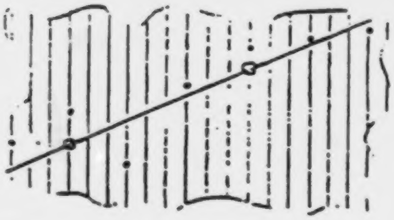


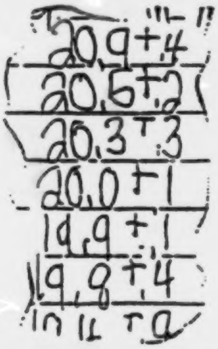


EDUCATIONAL TESTING SERVICE, ETS and the ETS logo are registered trademarks of Educational Testing Service.
PACKETS™ is a trademark of Educational Testing Service. Copyright © 1993 by Educational Testing Service.
All rights reserved.

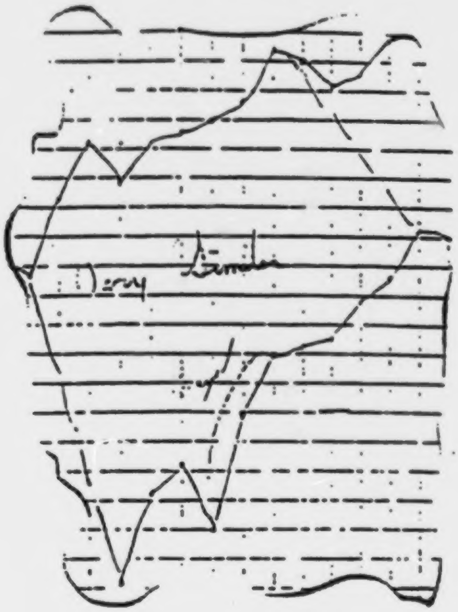
Numerical Patterns vs Graphical Patterns

How do people think about data? Sometimes people think in terms of numerical data, such as numbers, sequences of numbers (e.g., lists or tables) or composites of numbers (e.g., sums or averages). Other times they think in terms of visual data, such as points, sequences of points (e.g., patterns or graphs) and composites of points (e.g., slopes).



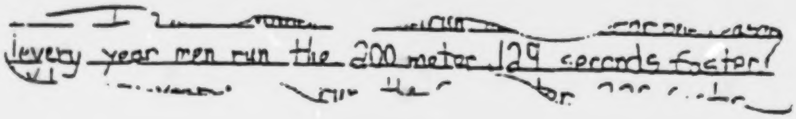


20.9 +.4
20.6 +.2
20.3 +.3
20.0 +.1
19.9 +.1
19.8 +.4
19.4 +.9

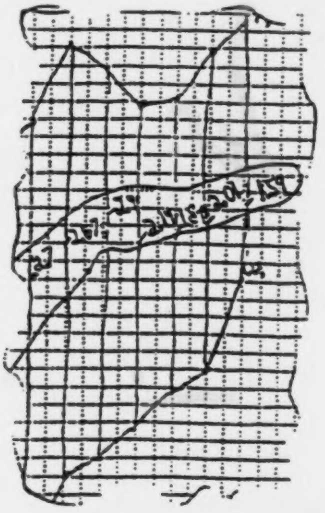


Independent Data vs Comparative Data

How do people think about more than one set of data? Sometimes people respond to men's and women's data separately, project into the future, then compare the two. Other times they consider differences between the two sets of data, then project the differences.



every year men run the 200 meter .129 seconds faster



EDUCATIONAL TESTING SERVICE, ETS and the ETS logo are registered trademarks of Educational Testing Service.
PACKETS™ is a trademark of Educational Testing Service. Copyright © 1993 by Educational Testing Service.
All rights reserved.

Sample Student Product
Group One

The Fast Track 4-17-92

I think women will outrun men soon. For one reason every year men run the 200 meter .129 seconds faster. Yet every year women run the 200 meter .326 faster. That means every year women gain .197 seconds every year. In around 20 years women will surpass men.

Avg. + .129 Avg. + .1288 100 years x.1289

Gain	Mens Time In Sec		Speed in mph		12.89
'88	19.75	+.05	22.6	+.1	
'84	19.80	+.39	22.5	+4	19.75
'80	20.19	+.04	22.1	0	1.289
'76	20.23	-.23	22.1	-.2	6.56
'72	20.00	-.17	22.3	-.2	
'68	19.83	+.47	22.5	+.5	25.8
'64	20.3	+.2	22.0	+.2	
'60	20.5	+.1	21.9	+.1	
'56	20.6	+.1	21.7	+.1	
'52	20.7	+.4	21.6	+.4	
'48	21.1	-.4	21.2	-.4	
'36	20.7	+.5	21.6	+.5	
'32	21.2	+.6	21.1	+.6	
'28	21.8	-.2	20.5	-.2	
'24	21.6	+.4	20.7	+.4	
'20	22.0	-.3	20.3	-.3	
'12	21.7	+.9	20.6	+.8	
'08	22.6	-1.0	19.8	-.9	
'04	21.6	+.6	20.7	+.6	
'00	22.2		20.1		

	Men Time In Sec		
ter '20		+.129	
'16		+.129	
'12		+.129	
'08		+.129	
'04		+.129	
'00	19.	+.129	
'96	19.51	+.129	
'92 4	19.68	+.129	+.3

.326 .35

	Women Time in sec	Speed in mph
1988	21.34 +.47	20.9 +.4
1984	21.81 +.22	20.5 +.2
1980	22.03 +.24	20.3 +.3
1976	22.37 +.03	20.0 +.1
1972	22.4 +.1	19.9 +.1
1968	22.5 +.5	19.8 +.4
1964	23.0 -1.0	19.4 +.8
1960	24.0 -.6	18.6 -.5
1956	23.4 +.3	19.1 +.3
1952	23.7 +.7	18.8 -.5
1948	24.4	18.3

BEST COPY AVAILABLE

EDUCATIONAL TESTING SERVICE, ETS and the ETS logo are registered trademarks of Educational Testing Service.
PACKETS™ is a trademark of Educational Testing Service. Copyright © 1993 by Educational Testing Service.

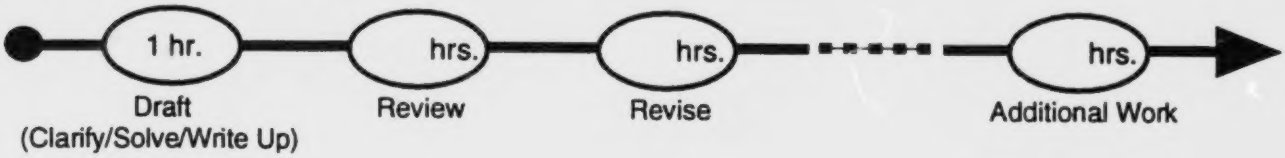
Sample Product Cover Sheet

(Conditions under which work was produced)

Who contributed to this work? What is their background?
(e.g., age, grade level, profession, etc.)

Four eighth grade students in an algebra preparation class.

How much time was spent on the activity?



1 hr. Draft (Clarify/Solve/Write Up)

hrs. Review

hrs. Revise

hrs. Additional Work

What resources were used?
(e.g., reference books, calculators, etc.)

Calculators

Additional Comments:

Copyright © 1993 by Educational Testing Service. All rights reserved.

Instructional Feedback Option

Fast Track

Student Group One

Dear Students,

Thank you for taking the time to make predictions for my article, *The Fast Track*. However, the information you sent is not complete enough for the basis of an article.

I like the argument that you gave on the first page, but wasn't completely sure that I understood where each of the numbers came from. (Where did 0.129 come from? Have you confused years with Olympics? Remember, Olympics occur every four years!) In the tables, it looks as if you calculated differences and then averaged them. Unfortunately, I need more than just the time when women will surpass men. I need predictions for men and women for the next 50 Olympics. Your predictions for the next eight Olympics are interesting, but not quite enough for me to use.

To develop this work further, try to project performance for 50 Olympics. The data need to be more readable, too. It would help if information was labelled and some explanation was given.

You have made a good start on this project. However, I need clearer and more complete information if I am to use it for my article.

Thanks,

The Editor

Copyright © 1993 by Educational Testing Service. All rights reserved.

Dimensions of Math PACKETS™ Interpretation Framework

	Mathematical Content in Response	Utility of Response
Description	What math was used?	What purpose(s) does the product serve?
Evaluation	How well was the math used?	How useful is the product for the purpose(s) stated in the activity?

BEST COPY AVAILABLE

Copyright © 1993 by Educational Testing Service. All rights reserved.

PACKETS™ Assessment
Steps for the Teacher

GATHER PROCESS INFORMATION
(Mathematical and Social)

- Observe students during the activity.
- Listen to and question students during their presentations.
- Lead/participate in classroom discussions.

DESCRIBE THE PRODUCT (Ways of Thinking)

- Analyze the mathematics used in the product.

EVALUATE THE PRODUCT (Windows on Quality)

- Assess the utility of the product from the point of view of the audience.
- Describe the strengths and weaknesses of the product.
- Suggest next steps to the student (i.e., the extent of revisions needed to make the product acceptable).

Copyright © 1993 by Educational Testing Service. All rights reserved.

WINDOWS ON QUALITY OF THE PRODUCT

The overarching consideration when assessing a piece of work should be how well it accomplishes the expressed purpose of the activity, the task at hand. One should look at the approach used and the results obtained, in relation to this purpose. To do this, one should keep in mind what students were asked to do, for whom, and for what purpose.

How well does the product accomplish the purpose of the activity?

(Does the product contain everything that was asked for or a substitute that meets the expressed needs?)

How well is the product supported by appropriate mathematics?

How appropriately and effectively are the mathematical concepts used?
How appropriately and effectively are the mathematical operations and procedures used?
How appropriately and effectively are the mathematical representations used?
How skillfully are the mathematics used?

How understandable is the product to the intended audience?

To what extent is the product –

- clear
- consistent
- complete
- coherent
- well-organized
- aesthetic?

How reasonable is the product for the real-world situation?

How logical are the solution paths from "the givens" in the problem to the final product?
How is the solution supported, justified, or explained?
How much and what kind of information was used and/or generated?
Was information omitted, ignored, distorted, or invented? If so, how?

Is there something special in this product?

Connections (within mathematics, across disciplines, or to the real-world)
Awareness of assumptions, sources of error, or limitations
Recognition of alternative approaches
Extensions and generalizations
Uniqueness
Anything else

Copyright © 1993 by Educational Testing Service. All rights reserved.

Performance Level Guides for Open-Response Common Items
Grade 8

Kentucky Department of Education

A major part of education reform in Kentucky is the Kentucky Instructional Results Information System (KIRIS). The annual assessment, at grades 4, 8, and 12, has three parts: multiple-choice and short-essay questions; performance tasks that call for students to work together in groups or individually to solve simulated, real-life problems; and portfolios that present each student's best work collected throughout the school year.

As part of the 1991-92 KIRIS Assessment, each eighth grader took three open-ended questions in reading, mathematics, science and social studies. The science and social studies questions, together with examples of student responses at the distinguished, proficient, apprentice, and novice performance levels, are reproduced here.

For more information on Kentucky's assessment program, contact:

Mr. Edward Reidy
Kentucky Department of Education
500 Mero Street
Capitol Plaza Tower
Frankfort, KY 40601

These materials were reproduced with the permission of the Kentucky State Education Department.

World-Class Standards...

for

World-Class Kids.

In Kentucky, we just expect more!

Performance Level Guides
for
Open-Response Common Items

GRADE 8

Kentucky Department of Education
Thomas C. Boysen, Commissioner

BEST COPY AVAILABLE

INTRODUCTION

As part of the 1991-92 KIRIS Assessment, each eighth grader took three open-ended questions in reading, mathematics, science and social studies. Those questions, together with examples of student responses at the distinguished, proficient, apprentice, and novice performance levels, are provided in this booklet. In addition to the open-ended questions in each of the four content areas, students were required to compile a writing assessment portfolio to demonstrate their proficiency in writing. The table of contents for the portfolio, along with examples of student writing at each performance level, are also included.

Grade 8

SCIENCE OPEN-RESPONSE COMMON ITEMS

Open-response 2:

How would life and the conditions on earth be different if all bacteria and fungi became extinct? Explain the changes that might occur and give as much detail as possible.

Open-response 3:

The table below shows the information a researcher has gathered about the students in the seventh grade in a school. Use this information to answer the questions that follow the table.

Name	Sex	Age	Best Subject	Number of Brothers & Sisters	Grade in Reading
Adams	M	12	M	2	80
Archer	M	12	R	0	85
Brown	M	13	M	2	76
Burton	F	12	R	2	84
Carrera	F	12	M	0	87
Davenport	M	13	S	1	85
Fenwick	M	12	M	0	79
Franklin	M	12	R	2	77
Garvey	M	12	M	1	81
Harris	F	12	S	3	70
Kelley	M	12	S	0	83
LaFontaine	M	12	M	1	80
Miranda	F	12	H	1	76
Moore	F	13	S	1	82
Peterson	M	12	M	1	86
Potvin	F	12	H	3	80
Sabin	M	13	S	0	90
Smith	F	12	M	2	83
Washburn	F	12	R	4	72
Weinberg	M	13	H	3	75
Wilson	F	13	M	2	79

Sex: M=Male F=Female
Best Subject: H=History, M=Math, R=Reading, S=Science

There are five groups of students based on the number of brothers and sisters each student has. Compute an average reading score for each group. Show your work.

- Make a graph of your results.
- What conclusion can you draw from your results?

Open-response 4:

Katie believes that students who do between 4 and 10 hours of homework per week make better grades than students who do not do homework or who do more than 10 hours of homework per week. To test this hypothesis, she is writing a survey that she will give to students at her school.

- What questions should Katie include in her survey?
- Describe the scientific procedure Katie should use.
- Describe what Katie should do with the responses to her survey to find if her hypothesis is correct.

— DISTINGUISHED LEVEL —
Student Response Samples*

The student at the distinguished level presents an articulate, thorough, and complete response to each question, using concise and precise analyses. The response clearly shows an understanding of scientific thought processes and procedures, the application of the same, and an extension beyond the expected achievement for a student at this grade level.

OPEN-RESPONSE 2

Though extremely minute, bacteria and fungi have an enormous role in life on Earth. Without them, many changes would occur. Bacteria break down dead organic matter so it will decay. If bacteria and fungi became extinct, the dead animals would never decompose. Instead, they would keep accumulating until there would be no room for humans to live. This would also cause serious environmental problems. The build-up of the dead organisms would create poor living conditions, not to mention the sickening smell. Decayed organic matter fertilizes the soil so that plants can grow. If the bacteria did not cause decay, the plants could not grow. If plants did not grow, humans ould be left without a food source. Food chains of every kind would be upset without the help of bacteria and fungi. Also, oxygen could not be produced if there were no plants. Some bacteria change the nitrogen in the air to a form that plants can use. Again, the lack of bacteria would be a hinderance to plant growth. Fungi and bacteria also cause and spread diseases. With the extinction of bacteria and fungi, many common diseases today may become extinct also. Though, this would seem like a good change, it could also be a problem. If no animal or human contracted a disease, the world would become overpopulated. In turn, a food shortage would occur. Life would be even more difficult without the diseases. Therefore, this world would certainly be harmed by the extinction of bacteria and fungi.

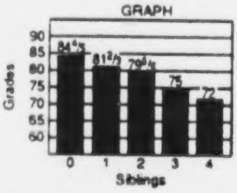
- in-depth, multifaceted, creative response
- analyzes both positive and negative aspects in detail
- focuses on the global aspects of this scenario
- includes the effects on the physical environment

OPEN-RESPONSE 3

My conclusion is, the less siblings you have the better in reading you are. You will have higher grades if you don't have any or very few brothers and sisters.

0 – 85 87 79 83 90
1 – 85 81 80 76 82 86
2 – 80 76 84 77 83 79
3 – 70 80 75
4 – 72

450 −26 424
5)424 84 4/5
40
24
80 ×6 480
6)490 81 2/3
400 +79
6)479 79 5/6
42
59
150 +75 225
3)225 75
21
15
1)72 72



- all necessary components are present and correct

OPEN-RESPONSE 4

Katie should ask [1] On the average, how many hours per week do you spend doing homework? [2] What are your grade point averages in Language Arts, Math, History, and Science? [3] What do you think influences your ability to do well? [4] Do you have any problems at home or school that could be influencing your grades? Katie should give this survey to a large number of students in all grades and classes and ask for their assistance in her experiment. She should take the information they give her and compile this data. The she should make a conclusion by discovering, by doing an average according to grades and how many hours of study, which group of students does the best and why. She could share her responses with a teacher to test her hypothesis and get a second opinion.

- all components are well-defined and articulated
- raises creative, valid questions beyond the expected
- attempts to verify results by sharing information with a teacher

---

* Wherever typed student responses appear, student errors have not been corrected.

— PROFICIENT LEVEL —
Student Response Samples

The student who has achieved a proficient level of performance has completed most of the important aspects of each open-response question, and is able to articulate the information required. While the overall quality of the responses displays proficiency, the student lacks the clarity and creativity present in the distinguished student.

OPEN-RESPONSE 2

> *If anything on earth dies out it will mess up the food chain and the land. If we no longer had fungi, and bacteria the dead organisms wouldn't have anything to eat. they'd have to find something new or die. chain reactions would go on for the entire food chain. People use a lot of fungi & bacteria in medicines so we would have to find something new too.*

- recognizes the global significance of bacteria and fungi
- demonstrates adequate analysis of the question
- lacks detail and clarity

OPEN-RESPONSE 3

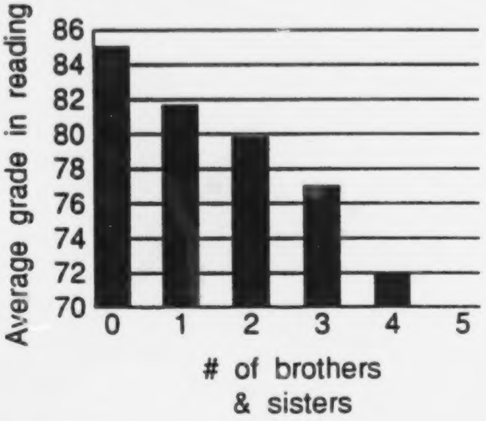
0	1	2	3	4
85	85	80	70	72
87	81	76	80	
79	80	84	75	
83	76	77	225	
90	82	83	75	
424	86	79		
$84\frac{4}{5}$	490	479		
	$81\frac{2}{3}$	$79\frac{5}{6}$		

5)424
40
24
20
4

6)490 $81\frac{4}{6}$
48
10

6)479
42
59
54
5

3)225
21
15



- shows correct averaging of the reading scores
- graph is well done and complete
- student has failed to generate a conclusion

OPEN-RESPONSE 4

> *Katie should ask these questions: "How many hours of homework do you do per week?" and "What grade do you receive?". First, Katie should test her hypothesis by conducting the survey. She should gather information from several people in order to get an accurate conclusion. Then, Katie must analize her results and draw a conclusion. Katie could make a graph of the survey to see if her hypothesis was correct. The graph would enable her to easily see the results.*

- two of the three crucial components answered adequately
- third component is sketchy and vague

— APPRENTICE LEVEL —
Student Response Samples

At this level, the student has completed some important portions of the open-response questions. There are significant gaps in the student's conceptual understanding.

This student has begun to grasp some important aspects of scientific processes and thinking patterns. The apprentice student lacks large and critical pieces of the good response but has begun to display some fundamental knowledge of portions of the questions.

OPEN-RESPONSE 2

> *Well, for one thing we would not have some medicines. For exp. penicillin is a bacteria medicine. Made mostly of molds. Even food is made from fungi and bacteria. So, we do need these things in life.*

- recognizes some important concepts
- lists two disadvantages to the loss of bacteria/fungi
- fails to list any advantages
- lacks any in-depth discussion
- fails to recognize global significance

OPEN-RESPONSE 3

> *The people with 0 siblings have better grades. Maybe cause they don't have to worry or fight with them. The people with a lot of siblings have bad grades probably cause they can't concentrate on studies.*
>
> *0 – 84.8%*
> *1 – 81.66%*
> *2 – 79.83%*
> *3 – 75%*
> *4 – 72%*

- misses crucial component by not making any attempt to graph the reading averages

OPEN-RESPONSE 4

> Questions: *How many hours do you study a night.*
> *Do you study before a test.*
> *What are your grades.*
> Procedure: *Gather as much data as possible and make a chart.*
> Response: *Give it to her teacher.*

- one of the three crucial components of this question has been adequately addressed
- understands the appropriate questions to ask to gather information
- does not understand how to conduct or analyze a scientific study

— NOVICE LEVEL —
Student Response Samples

A student at the novice level of performance does not grasp the question and has not responded to it in a meaningful way.

Responses indicate little or no understanding of the questions. Many responses are merely restatements of the question, and often do not make sense. Some students at the novice level will simply write nonsense when facing a question that they don't understand.

OPEN-RESPONSE 2

> *There would be know bacteria and dirt around, & I think that the world would be a whole lot cleanier and different.*

- fails to process question in a meaningful way
- reference to "dirt" is not appropriate for question

OPEN-RESPONSE 3

> *80, 76, 84, 85, 77, 81, 80, 76, 82, 86, = 807*
> *807 is the average*

- does not understand how to compute an average
- has not attempted to construct a graph
- has not drawn any conclusion

OPEN-RESPONSE 4

> *What do they study. After Katie got all the answers she should try to do them to see if it really works.*

- does not know appropriate questions to ask
- does not understand how to conduct a meaningful scientific study
- unable to respond in a meaningful way

Grade 8

SOCIAL STUDIES OPEN-RESPONSE COMMON ITEMS

Open-response 2:

In the United States, people are sometimes treated unfairly because of the group to which they belong.

- Identify a group that has been treated unfairly because of sex, national origin, religion or race.
- Describe several ways in which this group has been treated unfairly throughout United States history.
- Also describe several ways in which people have tried to correct these problems.

Open-response 3:

New York City, as of 1990, had a population of 7,322,564. Shelbyville, Kentucky, at the same time, had a population of 6,238.

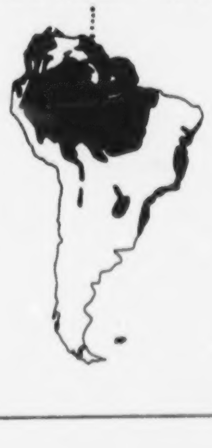
- What are SEVERAL opportunities people from a large urban center would claim they have that people from rural areas do not have?
- How would people from the rural areas probably argue against those claims?
- What are some problems that people from the two types of communities would have in common?

Open-response 4:

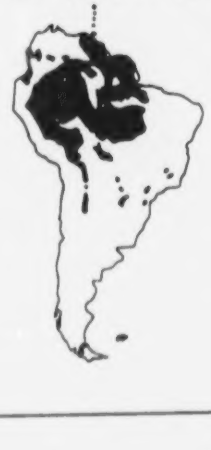
The shaded areas on the maps below indicate the extent of rain forests at two different times in the last 50 years.

- Describe several changes that have taken place in the rain forests that explain the differences between the maps.
- Why have these changes been occurring?

Extent of Tropical
Rain Forest – 1940



Extent of Tropical
Rain Forest – 1988



— DISTINGUISHED LEVEL —
Student Response Samples*

Students at this level demonstrate the ability to make interpretations, draw conclusions, summarize, analyze, and evaluate information, and provide logical arguments in support of a position based on their understanding of social studies knowledge and concepts. They can clearly communicate ideas and often "go beyond" what is asked for in a question by incorporating as many relevant ideas, information, and examples as possible.

OPEN-RESPONSE 2

> *We should take a lesson from the Japanese. They have risen from the ashes of WWII, and have become a world power. But instead, Americans choose to insult them.*
>
> *The Japanese have been discriminated against because of WWII and Pearl Harbor. They have been riduculed because they are buying American companies. Most recently, we have had a name-calling contest with them since their prime minister said that American workers were "lazy and illiterate."*
>
> *Unfortunately, we have not done much to combat discrimination of the Japanese. But if we don't like what they are doing by buying America, we must do something about it, and not by badmouthing them.*

- discusses group that has been discriminated against as well as the ways discrimination has occurred
- gives one possible solution for solving problem of discrimination

OPEN-RESPONSE 3

> *People who live in urban areas have distinct advantages over rural areas in some cases. They have ready access to jobs, banks, theatre, and almost anything else. There are more jobs available, and they have better education.*
>
> *However, in small towns, you know just about everyone, there is less crime, and your voice can be heard more in politics.*
>
> *Unfortunately, we share some of the same problems: drug abuse, child abuse, kids who don't want to learn, and people who don't like each other. These are problems you will face anywhere you go.*

- discusses advantages and disadvantages of both urban and rural areas
- gives several problems that are common to both urban and rural areas

OPEN-RESPONSE 4

> *Every day, we lose hundreds of acres of tropical rain forest. Over 50 years, we have wiped out a lot of animals and their habitat. The land on the maps has been cleared. There are several reasons why. First is for farming. But, the soil in the rain forest is poor, and it can only support crops for a few years. Farmers must clear more land. Second is for building homes. Third, many forests are being cut down for their expensive woods, like ebony, teak, and mahogany. We must stop clearing rain forests. We are killing animals and robbing ourselves of a wonderful treasure.*

- gives accurate description of maps
- discusses reasons for the disappearance of the rain forests
- discusses problems that deforestation creates

---

* Wherever typed student responses appear, student errors have not been corrected.

— PROFICIENT LEVEL —
Student Response Samples

Proficient students can demonstrate an understanding of social studies concepts and/or issues and generate a coherent discussion based on them. Students may be able to draw plausible conclusions in light of their social studies knowledge, but may not be able to provide adequate support for their positions.

OPEN-RESPONSE 2

> *The Native Americans have always been treated wrongly. Back during those first years of our country, settlers took their land from them, destroyed their camps, killed the Indians' source of food and waged an unrightful war against them. The Native Americans, once roaming the whole nation, were forced to live on small reservations. At the beginning, there were men that warned Spain of the wrongs committed against Indian. Since then, there have been people protesting against it, even during the settlers' days. There were never enough, though and that should be our greatest shame.*

- indicates an understanding of a group that has been treated harshly and provides the reasons
- gives historical basis for the answer, but is minimal in its breadth and scope

OPEN-RESPONSE 3

> *There would be better schooling, more cultural centers, more opportunities for people to become involved with activities. This would all happen because there are more people, more needs, and more tax money. Rural people may feel that since there are less students, teachers may spend more time with the children. They may feel that their cultural centers may be fewer in number, but better in content. They may feel that the opportunities for the activities would be better handled because there are less people to handle. Both communities would have to worry about crime, pollution, and homeless people.*

- indicates a knowledge of advantages and disadvantages associated with urban and rural living, with some inaccuracies

OPEN-RESPONSE 4

> *The rain forests have definitely been shrinking. Every year thousands of acres are being destroyed. Trees are cut down and land cleared and with it all, habitats for as yet unknown species. Plants that could produce life-saving medicines are being killed. The rain forests are cleared because of many reasons. The inhabitants of S.A. need land to grow crops, lumber to sell, and animals to export for pets. It all comes down to the basic role of money.*

- shows an understanding of, reasons for, and effects of deforestation

— APPRENTICE LEVEL — Student Response Samples

A student at this level displays some understanding of social studies concepts. The student may clearly communicate some accurate knowledge, but a thorough understanding is not apparent. An apprentice level student is beginning to incorporate facts with processing skills to develop a clear response.

OPEN-RESPONSE 2

> *I think mainly in our history blacks have been treated unfairly. During colonial times many blacks were being used as slaves. During the 40's, 50's, and 60's many blacks were put under public humiliation under segregation and the lack of basic rights that the whites had. During the years black leaders like Martin Luther King Jr and Malcolm X have stood up for black right and are probably the main reasons why blacks have the same rights today as everyone else.*

- indicates some insight as to ways in which a group of people have been treated unfairly
- provides minimal description of how people have tried to correct the problem
- implies that problems have been resolved

OPEN-RESPONSE 3

> *People from a large urban center would claim that they have easier access to things like shopping centers and food markets. They would say that they have a special closeness with their neighbors and that they could trust them. People from a rural area would propably say that in a city they have more crimes and it's dangerous to walk the streets at night, but in the country is safe and quiet. They might have fire and theft problems alike.*

- generally lacks explanation and insight
- gives some ways in which urban and rural areas are both alike and different

OPEN-RESPONSE 4

> *There are many changes that have occured in the rain forest. Everyday men are bulldozing and burn down rainforests just for lan to raise cattle on. Many companies are doing the same thing but using the land for apartment houses, shopping centers, and other odd sources of buildings. These changes mainly have been occuring because the population is steadily increasing in South America. With the rise in population they need more land to live on.*

- indicates some understanding of why the rain forests are shrinking
- gives possible reasons but little explanation

— NOVICE LEVEL — Student Response Samples

A student at this level is unable to clearly communicate important ideas that would indicate an understanding of social studies concepts. Discussion of social studies issues at this level may be purely recall of fact with no strategy for processing the information coherently.

OPEN-RESPONSE 2

> *Sex is not all unfairly and pluse is you think sex is unfairly. It is to alot of young boys and girls. They shouldn't be talking about sex anyways.*

- fails to identify a group that has been treated unfairly
- shows minimal understanding of the question

OPEN-RESPONSE 3

> *They both have people and schools and jobs maybe the school work is I don't know. i go to Dayton. New York City has more people.*

- lacks any attempt to compare and contrast characteristics of urban and rural areas

OPEN-RESPONSE 4

> *In 1940 there is a lot of rain forests that In 1988 has less rain forest*

- makes a statement in reference to the change that has taken place in South American rain forest regions
- no attempt to cite causes or effects of rain forest loss

Advanced Placement Program

1992 Mathematics: Free-Response Scoring Guide and Sample Student Answers

Calculus AB – Calculus BC

"Message to Teachers," sample student responses, scoring guides for question 1 of the 1992 AP Calculus AB Examination, and "Reminders for Secondary School Teachers" are reproduced here.

The Advanced Placement Program is a cooperative educational endeavor sponsored by the College Board and administered by Educational Testing Service. The program serves three groups: students who wish to pursue college-level studies while still in high school, high schools that wish to offer these opportunities, and colleges that wish to encourage and recognize such achievement. The Program provides materials describing college-level courses to participating high schools and the results of examinations based on these courses to the colleges of the students' choice. Participating colleges, in turn, grant credit and/or appropriate placement to students who demonstrate qualifying performance on the examinations.

Except for Studio Art — a portfolio assessment — the AP Examinations are approximately 50 percent free-response (essays, problems, programs, taped performances, etc.). Course descriptions, teachers' guides, released examinations, free-response guides, student guides, and other curricular materials are available in art, biology, chemistry, computer science, economics, English, French, German, government and politics, history, Latin, mathematics, music, physics, psychology and Spanish. Throughout the country, in all of these disciplines, teacher development workshops are available year-round and week-long institutes are available in the summer.

For more information, contact:

Advanced Placement Program
Mail Stop 85-D
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

These materials were reproduced with the permission of the College Board and Educational Testing Service.

1992 AP Mathematics: Free-Response Scoring Guide and Sample Student Answers

Calculus AB — Calculus BC

Dear Teacher:

This *Guide* is an attempt to remove the mystery of how AP Mathematics free-response questions are scored.

As you probably know, AP Examinations are created afresh each year by faculty development committees and Educational Testing Service (ETS) test development specialists. All the committee members teach calculus in colleges, universities, and secondary schools. The multiple-choice questions are pretested in college calculus classes; they are then analyzed and scored so that the committee is able to judge their suitability for inclusion in the final examinations that AP students take each May.

The scoring of the free-response questions takes place during June at the AP Reading. Before the faculty consultants — college and AP teachers from around the country — gather at the Reading, the chief faculty consultant, Professor Raymond J. Cannon of Baylor University, together with the two examination leaders, prepared the first draft of the free-response scoring guides at Clemson University in South Carolina before the arrival of the 40 table leaders. The scoring guides were refined after the table leaders scored the full range of sample responses. Using the agreed upon scoring guides and sample papers, the table leaders trained the 322 faculty consultants who arrived at Clemson on June 13. After the training, using the samples, faculty consultants were ready to score actual student responses. Throughout the Reading, the table leaders continuously monitored the faculty consultants, checking their scores and discussing particular student responses with them. Over the course of six days, 77,508 Calculus AB papers and 15,639 Calculus BC papers were read.

To arrive at a total grade, the scores for each question given by faculty consultants were weighted and combined with the multiple-choice scores. Armed with statistical data about student performance, the apparent ability of the student group, and past score distributions, the chief faculty consultant set the grade "cut points" after consulting with ETS and College Board staff.

This *Guide* illustrates the range and quality of student work. For each free-response question from the 1992 AP Mathematics Examinations, the chief faculty consultant has provided a general comment and an observation about the performance of the candidates. Next you will find the scoring guides. Probably most useful in your work will be the actual student responses and the chief faculty consultant's comments on those responses explaining why they merited the scores they received. National grade distributions follow, along with information about how to interpret them.

We hope this publication will be a valuable aid to you in your teaching.

Wade Curry

Wade Curry, Director
Advanced Placement Program

I.
THE 1992 AP CALCULUS AB EXAMINATION

Free-Response Questions, Scoring Guides, and Sample Student Answers

Question 1

This opening problem requires some basic analysis of a polynomial graph. Students should have answered Part (a) by considering the sign of the first derivative, noting that it remains negative on both sides of the critical point at 0. Part (b) was to be answered by considering the sign of the second derivative. In Part (c), students had to indicate that horizontal tangent lines occur at all three critical points, including the "shelf" point at $x = 0$.

A favorite technique for analyzing the sign of $f'(x)$ and $f''(x)$ is through a "sign chart." Students were required to have labeled such charts clearly, and to indicate what information was being used to draw their conclusions.

1. Let f be the function defined by $f(x) = 3x^5 - 5x^3 + 2$.

   (a) On what intervals is f increasing?

   (b) On what intervals is the graph of f concave upward?

   (c) Write the equation of each horizontal tangent line to the graph of f.

Solution

Scoring Scale

Points (See Footnote)

(a) $f'(x) = 15x^4 - 15x^2 = 15x^2(x^2 - 1)$

$$\begin{array}{lccccccc} \text{Sign of } f' & + & & - & & - & & + \\ \hline & & -1 & & 0 & & 1 & \end{array}$$

Answer: f is increasing on the intervals $(-\infty, -1]$ and $[1, \infty)$

$$3\begin{cases} 1:\ f'(x) \\ 1:\ \text{Analyzes sign of } f'(x) \text{ or} \\ \quad\ \text{explicitly sets student's} \\ \quad\ f'(x) > 0 \\ 1:\ \text{Answer} \end{cases}$$

(b) $f''(x) = 60x^3 - 30x = 30(2x^2 - 1)$

$$\begin{array}{lccccccc} \text{Sign of } f'' & - & & + & & - & & + \\ \hline & & -\frac{1}{\sqrt{2}} & & 0 & & \frac{1}{\sqrt{2}} & \end{array}$$

Answer: on $\left(-\frac{1}{\sqrt{2}}, 0\right)$

and on $\left(\frac{1}{\sqrt{2}}, \infty\right)$

$$3\begin{cases} 1:\ f''(x) \\ 1:\ \text{Analyzes sign of } f''(x) \text{ or} \\ \quad\ \text{explicitly sets student's} \\ \quad\ f''(x) > 0 \\ 1:\ \text{Answer} \end{cases}$$

Footnote: Whenever points are deducted for specific errors as denoted in < >, students cannot lose more points than the number of points indicated for that part in the scoring scale.

Copyright © 1992 by Educational Testing Service. All rights reserved.
Princeton, N.J. 08541

(c) $f'(x) = 0$ when $x = -1, 0, 1$

$x = -1 \Rightarrow f(x) = 4;\ y = 4$

$f(0) = 2;\ y = 2$

$f(1) = 0;\ y = 0$

3
- 1: Solves student's $f'(x) = 0$
  - $< -1 >$ fewer than 3 solutions
- 1: One answer
- 1: All other consistent answers (must be at least one)

Note: For "answer only," maximum 2 of 3 if no f'

Student Response

1. Let f be the function defined by $f(x) = 3x^5 - 5x^3 + 2$.

(a) On what intervals is f increasing?

$f'(x) = 15x^4 - 15x^2$

$f'(x) = 15x^2(x^2 - 1)$

critical values

$15x^2(x^2-1) = 0$

$x = 0, x = 1, x = -1$

	$f'(x)$	$f(x)$ is	
$(-\infty, -1)$	+	increasing	$f'(-2) = 60(3)$
$(-1, 0)$	−	decreasing	$f'(-\frac{1}{2}) = \frac{15}{4}(-\frac{3}{4})$
$(0, 1)$	−	decreasing	$f'(\frac{1}{2}) = \frac{15}{4}(-\frac{3}{4})$
$(1, +\infty)$	+	increasing	$f'(2) = 60(3)$

$f(x)$ is increasing on $(-\infty, -1) \cup (1, +\infty)$

(b) On what intervals is the graph of *f* concave upward?

$f''(x) = 60x^2 - 30x$

$f''(x) = 30x(2x^2 - 1)$

Possible points of inflection

$f''(x) = 30x(2x^2 - 1) = 0$

$x = 0 \quad x = \frac{1}{\sqrt{2}} \quad x = \frac{-1}{\sqrt{2}}$

	$f''(x)$	$f(x)$ is concave	
$(-\infty, \frac{-1}{\sqrt{2}})$	−	downwards	$f''(-1) = -60 + 30 = -30$
$(\frac{-1}{\sqrt{2}}, 0)$	+	upwards	$f''(\frac{-1}{2}) = \frac{-60}{8} + 15 = \frac{15}{2}$
$(0, \frac{1}{\sqrt{2}})$	−	downwards	$f''(\frac{1}{2}) = \frac{60}{8} - 15 = \frac{-15}{2}$
$(\frac{1}{\sqrt{2}}, \infty)$	+	upwards	$f''(1) = 60 - 30 = 30$

$f(x)$ is concave upwards on $(\frac{-1}{\sqrt{2}}, 0) \cup (\frac{1}{\sqrt{2}}, +\infty)$

(c) Write the equation of each horizontal tangent line to the graph of *f*.

critical values at $x = 0$, $x = -1$, $x = 1$

$f(0) = 2 \quad f(-1) = 4 \quad f(1) = 0$

horizontal tangent lines at:

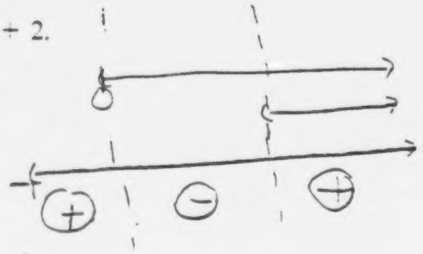
$y = 2$; $y = 4$, $y = 0$

Comment: In Parts (a) and (b), note how clearly this student communicates by first labeling the sign charts of *f'* and *f''* and then showing the corresponding conclusions about the behavior of *f*. Because of this clarity, faculty consultants were not confused when the student wrote the two "prime" signs for *f''* so close together that they may appear as one. The student is not penalized for the grammatical error in Part (c), using "at" rather than "are." Similarly, the student's answer to Part (a) is technically incorrect, since *f* is not increasing on the set that is the union of these two intervals. However, because the question asked for the intervals, the union sign was interpreted as "and on" and credit was awarded. The score for this solution is 9.

Student Response

1. Let f be the function defined by $f(x) = 3x^5 - 5x^3 + 2$.

(a) On what intervals is f increasing?



$f'(x) = 15x^4 - 15x^2$

$15x^2(x^2-1) > 0$

$15x^2 > 0$

$\boxed{x > 0}$ critical pnts

$x^2 > 1$ $-1, 0, 1$

$\boxed{x > \pm 1}$

$\boxed{f \text{ is increasing } [-1, 0] \text{ and } [1, \infty]}$

(b) On what intervals is the graph of f concave upward?

$f'(x) = 15x^4 - 15x^2$

$f''(x) = 60x^3 - 30x$ pnts of inflection

$30x(2x^2 - 1)$ $0, \pm\sqrt{1/2}$

$30x > 0$

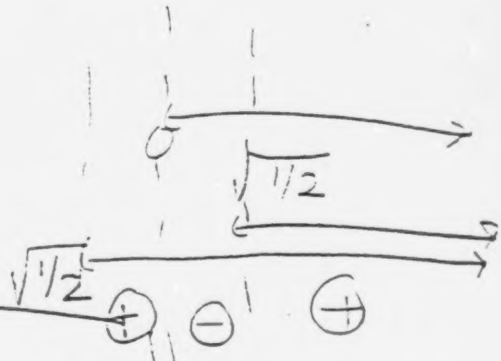
$x > 0$

$2x^2 - 1 > 0$

$2x^2 > 1$

$x^2 > 1/2$

$x > \pm\sqrt{1/2}$



$\boxed{f \text{ is concave up at } [-\sqrt{1/2}, 0] \text{ and } [\sqrt{1/2}, \infty]}$

(c) Write the equation of each horizontal tangent line to the graph of f.

$f(x) = 3x^5 - 5x^3 + 2$

$f'(x) = m = 15x^4 - 15x^2 = 0$

$15x^2(x^2-1) = 0$

~~$x = \pm 1$~~

~~$x = 0$~~

Comment: In Part (a), this student showed algebraically that $f'(x) > 0$, but failed to solve that inequality correctly. In Part (b), he or she managed, through some algebra that by itself appears suspicious, to solve the desired inequality correctly. Note how the student used sign charts without indicating what they mean, making it impossible to assign partial credit if the answer had been incorrect. In Part (c), the student missed the first of the three points available because the answers were crossed off. Historically, a penalty has been assigned for crossing off correct work, and here that penalty amounted to one point. In the future, crossed off work will not be scored. The score for this solution is 5.

Reminders for Secondary School Teachers

AP Examinations are designed to provide accurate assessments of achievement when the results are used properly. Any examination has limitations, however, especially when used for purposes other than those intended. Presented below are general and specific suggestions for teachers to aid in the use and interpretation of AP grades.

- AP Examinations are developed and evaluated independently of each other. They are linked only by common purpose, format, and method of reporting results. Therefore, comparisons should not be made between grades on different AP Examinations. An AP grade in one subject may not have the same meaning and interpretation as the same AP grade in another subject, just as national and college standards vary from one discipline to another.

- AP grades are not exactly comparable to college course grades. The AP Program conducts research studies every few years in each AP subject to ensure that the AP grading standards are comparable to those used in colleges with similar courses. In general, these studies indicate that an AP grade of 3 is approximately equal to a college course grade of B at many institutions. At some other institutions, an AP grade of 3 is more nearly comparable to a college course grade of C. These are only generalizations, however. The degree of comparability of an AP grade to a college course grade depends to a large extent on the particular college course used for comparison.

- The confidentiality of candidate grade reports should be recognized and maintained. All individuals who have access to AP grades should be aware of the confidential nature of the grades and agree to maintain their security. In addition, school districts and states should not release data about high school performance without the school's permission.

- AP Examinations are not designed as instruments for teacher or school evaluation. A large number of factors influence AP Exam performance in a particular course or school in any given year. As a result, differences in AP Exam performance should be carefully studied before attributing them to the teacher or school.

- Where evaluation of AP students, teachers, or courses is desired, local evaluation models should be developed. An important aspect of any evaluation model is the use of an appropriate method of comparison or frame of reference to account for yearly changes in student composition and ability, as well as local differences in resources, educational methods, and socioeconomic factors.

- The "Report to AP Teachers" can be a useful diagnostic tool in reviewing course results. The report identifies areas of strength and weakness for each AP course. This information may also help to guide your students in identifying their own strengths and weaknesses in preparation for future study.

- Many factors can influence course results. AP Exam performance may be due to the degree of agreement between your course and the course defined in the relevant AP *Course Description*, use of different instructional methods, differences in emphasis or preparation on particular parts of the examination (e.g., writing and organizational skills), differences in pre-AP curriculum, or differences in student background and preparation in comparison with the national group.

“Performance Assessment”

Education Research Consumer Guide,
Number 2, November 1992

This guide to performance assessment, produced by the U.S. Department of Education, defines the topic and provides some basic information including what the research says, the cost of performance assessment, and provides some examples of successful strategies and programs. People to contact for more information are also listed.

Education Research Consumer Guide is a new series designed for teachers, parents, and others interested in current education themes. Jacqueline Zimmermann is the editor. It is published by the Office of Research, Office of Educational Research and Improvement, U.S. Department of Education.

Education Research Consumer Guide

Number 2

November 1992

Performance Assessment

What is it? Performance assessment, also known as alternative or authentic assessment, is a form of testing that requires students to perform a task rather than select an answer from a ready-made list. For example, a student may be asked to explain historical events, generate scientific hypotheses, solve math problems, converse in a foreign language, or conduct research on an assigned topic. Experienced raters—either teachers or other trained staff—then judge the quality of the student's work based on an agreed-upon set of criteria. This new form of assessment is most widely used to directly assess writing ability based on text produced by students under test instructions.

How does it work? Following are some methods that have been used successfully to assess performance:

- **Open-ended or extended response exercises** are questions or other prompts that require students to explore a topic orally or in writing. Students might be asked to describe their observations from a science experiment, or present arguments an historic character would make concerning a particular proposition. For example, what would Abraham Lincoln argue about the causes of the Civil War?
- **Extended tasks** are assignments that require sustained attention in a single work area and are carried out over several hours or longer. Such tasks could include drafting, reviewing, and revising a poem; conducting and explaining the results of a science experiment on photosynthesis; or even painting a car in auto shop.
- **Portfolios** are selected collections of a variety of performance-based work. A portfolio might include a student's "best pieces" and the student's evaluation of the strengths and weaknesses of several pieces. The portfolio may also contain some "works in progress" that illustrate the improvements the student has made over time.

These methods, like all types of performance assessments, require that students actively develop their approaches to the task under defined conditions, knowing that their work will be evaluated according to agreed-upon standards. This requirement distinguishes performance assessment from other forms of testing.

Why try it? Because they require students to actively demonstrate what they know, performance assessments may be a more valid indicator of students' knowledge and abilities. There is a big difference between answering multiple choice questions on how to make an oral presentation and actually making an oral presentation.

More important, performance assessment can provide impetus for improving instruction, and increase students' understanding of what they need to know and be able to do. In preparing their students to work on a performance task, teachers describe what the task entails and the standards that will be used to evaluate performance. This requires a careful description of the elements of good performance, and allows students to judge their own work as they proceed.

What does the research say? *Active learning.* Research suggests that learning how and where information can be applied should be a central part of all curricular areas. Also, students exhibit greater interest and levels of learning when they are required to organize facts around major concepts and actively construct their own understanding of the concepts in a rich variety of contexts. Performance assessment requires students to structure and apply information, and thereby helps to engage students in this type of learning.

Curriculum-based testing. Performance assessments should be based on the curriculum rather than constructed by someone unfamiliar with the particular state, district or school curriculum. This allows the curriculum to "drive" the test, rather than be encumbered by testing requirements that disrupt instruction, as is often the case. Research shows that most teachers shape their teaching in a variety of ways to meet the requirements of tests. Primarily because of this impact of testing on instruction, many practitioners favor test reform and the new performance assessments.

Worthwhile tasks. Performance tasks should be "worth teaching to"; that is, the tasks need to present interesting possibilities for applying an array of curriculum-related knowledge and skills. The best performance tasks are inherently instructional, actively engaging students in worthwhile learning activities. Students may be encouraged by them to search out additional information or try different approaches, and in some situations, to work in teams.

What does it cost? These positive features of performance assessment come at a price. Performance assessment requires a greater expense of time, planning and thought from students and teachers. One teacher reports, "We can't just march through the curriculum anymore. It's hard. I spend more time planning and more time coaching. At first, my students just wanted to be told what to do. I had to help them to start thinking."

Users also need to pay close attention to technical and equity issues to ensure that the assessments are fair to all students. This is all the more important as there has been very little research and development on performance assessment in the environment of a high stakes accountability system, where administrative and resource decisions are affected by measures of student performance.

What are examples of successful strategies and programs?

- Charlotte Haguchi is a third- and fourth-grade teacher at Farmdale Elementary School in Los Angeles. Regarding assessment and instruction as inseparable aspects of teaching, Ms. Haguchi uses a wide array of assessment strategies to determine how well her students are doing and to make instructional decisions. She uses systematic rating procedures, keeps records of student performances on tasks, and actively involves students in keeping journals and evaluating their own work. Ms. Haguchi can be seen in action along with other experts and practitioners in the videotape *Alternatives for Measuring Performance* by NCREL and CRESST. (See Jeri Nowakowski and Ron Dietel, below.)
- William Symons is the superintendent of Alcoa City Schools in Alcoa, Tennessee. Seeking higher, more meaningful student standards through curriculum reform, Dr. Symons works with school staff and the community to create a new curriculum focused on standards and an assessment linked to the curriculum. Comments and advice from Dr. Symons and other practitioners and experts are available on the audiotape *Conversations About Authentic Assessment* by Appalachia Educational Laboratory. (See Helen Saunders, below.)
- Ross Brewer is the director of planning and policy development in the Vermont Department of Education. Vermont is assessing fourth- and eighth-grade students in writing and mathematics using three methods: a portfolio, a "best piece" from the portfolio, and a set of performance tasks. Other states that have been very active in developing and implementing performance assessments include: California, Arizona, Maryland, New York, Connecticut, and Kentucky. (See Ed Roeber and state officers, below.)

Where can I get more information?

W. Ross Brewer
Planning and Policy Department
Vermont Department of Education
Montpelier, VT 05602
(802)828–3135

Carolyn D. Byrne
Division of Educational Testing
New York State Education Department
Room 770 EBA
Albany, NY 12234
(518)474–5902

Dale Carlson
California Department of Education
721 Capitol Mall
Sacramento, CA 95814
(916)657–3011

Don Chambers
National Center for Research in Mathematical
Sciences Education
University of Wisconsin at Madison
1025 West Johnson Street
Madison, WI 53706
(608)263–4285

Ron Dietel
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA
145 Moore Hall
405 Hilgard Avenue
Los Angeles, CA 90024–1522
(310)206–1532

Steven Ferrara
Program Assessment Branch
Maryland Department of Education
200 West Baltimore Street
Baltimore, MD 21201
(410)333–2369

James Gilchrist
New Standards Project
Learning, Research and Development Center
3939 O'Hara Street
Pittsburg, PA 15260
(412) 624–8319

Paul Koehler
Arizona Department of Education
1535 West Jefferson
Phoenix, AZ 85007
(602)542–5754

Kate Maloy
National Research Center on Student Learning/LRDC
3939 O'Hara Street
Pittsburgh, PA 15260
(412)624–7457

Joe McDonald
Coalition of Essential Schools
Brown University
Box 1969
Providence, RI 02912
(401)863–3384

Jeri Nowakowski
North Central Regional Educational Laboratory (NCREL)
1900 Spring Road, Suite 300
Oak Brook, IL 60521
(708)571–4700

Edward Reidy
Office of Assessment and Accountability
Kentucky Department of Education
19th Floor Capital Plaza Tower
500 Mero Street
Frankfort, KY 40601
(502)564–4394

Douglas Rindone
Division of Research, Evaluation and Assessment
Connecticut Department of Education
Box 2219
Hartford, CT 06145
(203)566–1684

Ed Roeber
Council of Chief State School Officers
1 Massachusetts Avenue NW
Suite 700
Washington, DC 20001–1431
(202)336–7045

Larry Rudner
ERIC Clearinghouse/AIR
3333 K Street NW
Suite 300
Washington, DC 20007
(202)342–5060

Helen Saunders
Appalachia Educational Laboratory
1031 Quarrier Street
P.O. Box 1348
Charleston, WV 25325
(304)347–0400

by David Sweet

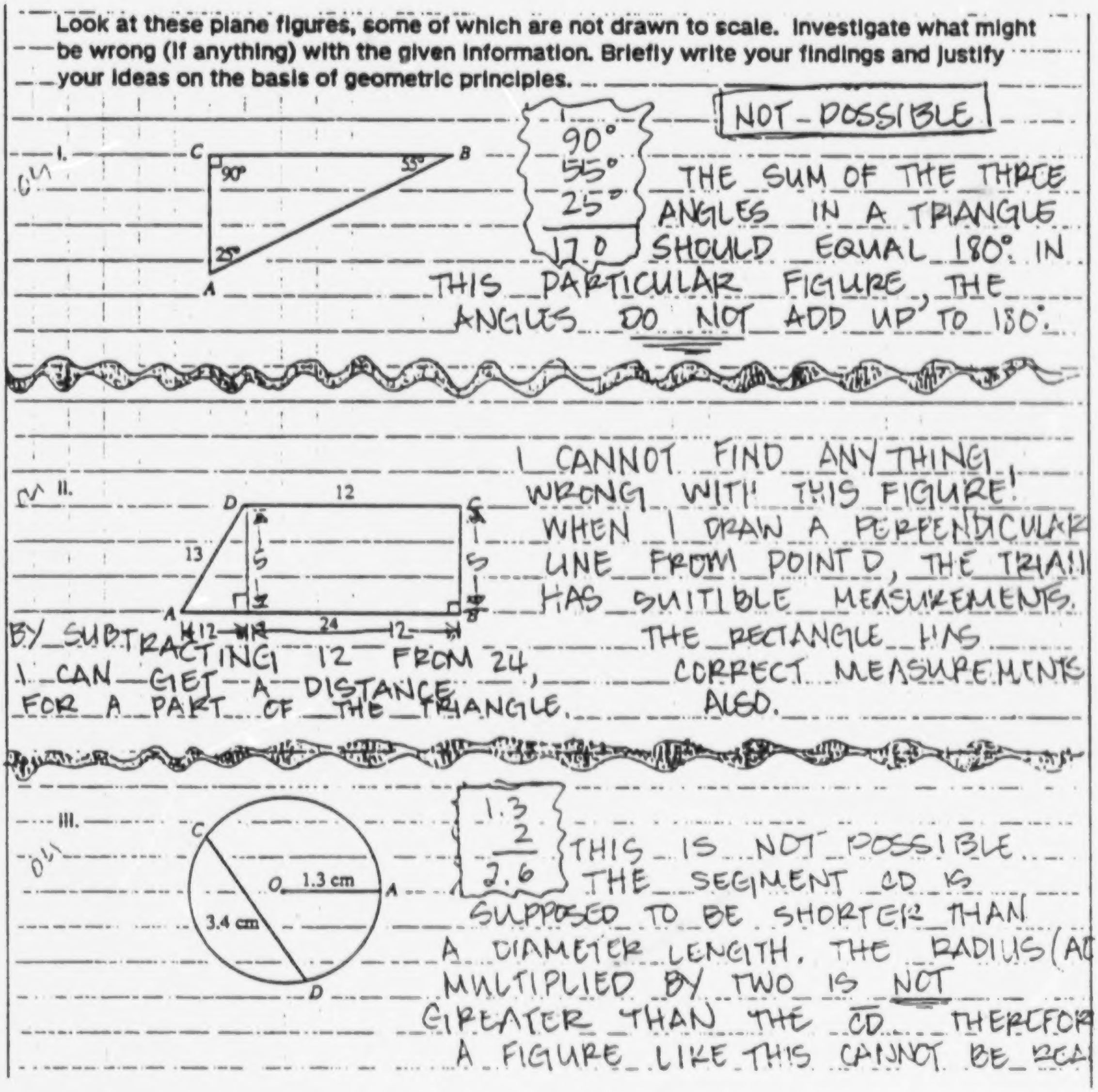
This is the second *Education Research CONSUMER GUIDE*—a new series published for teachers, parents, and others interested in current education themes.

OR 92–3056

ED/OERI 92-38

Editor: Jacquelyn Zimmermann

An Open-Ended Exercise in Mathematics: A Twelfth Grade Student's Performance


C
B
A
90°
55°
25°
D
12
13
5
5
12
24
12
O
1.3 cm
3.4 cm
C
A
D

Look at these plane figures, some of which are not drawn to scale. Investigate what might be wrong (if anything) with the given information. Briefly write your findings and justify your ideas on the basis of geometric principles.

I.

90°
55°
25°
170

NOT - POSSIBLE

THE SUM OF THE THREE ANGLES IN A TRIANGLE SHOULD EQUAL 180°. IN THIS PARTICULAR FIGURE, THE ANGLES DO NOT ADD UP TO 180°.

II.

I CANNOT FIND ANYTHING WRONG WITH THIS FIGURE! WHEN I DRAW A PERPENDICULAR LINE FROM POINT D, THE TRIA HAS SUITIBLE MEASUREMENTS.

BY SUBTRACTING 12 FROM 24, I CAN GET A DISTANCE FOR A PART OF THE TRIANGLE.

THE RECTANGLE HAS CORRECT MEASUREMENTS ALSO.

III.

1.3
2
2.6

THIS IS NOT POSSIBLE. THE SEGMENT CD IS SUPPOSED TO BE SHORTER THAN A DIAMETER LENGTH. THE RADIUS (A MULTIPLIED BY TWO IS NOT GREATER THAN THE $\overline{CD}$. THEREFOR A FIGURE LIKE THIS CANNOT BE REA

Reprinted, by permission, from *A Question of Thinking: A First Look at Students' Performance on Open-ended Questions in Mathematics*, copyright 1989, California Department of Education, P.O. Box 271, Sacramento, CA 95812–0271.

Education Research Consumer Guide is produced by the Office of Research, Office of Educational Research and Improvement (OERI) of the U.S. Department of Education.

Lamar Alexander, Secretary of Education • Diane Ravitch, Assistant Secretary, OERI

Francie Alexander, Deputy Assistant Secretary, OERI
Joseph C. Conaty, Acting Director, Office of Research

BEST COPY AVAILABLE

“A Look at a Middle School Portfolio”

Arts PROPEL: A Handbook for Visual Arts

Arts PROPEL is an approach to education that has evolved in the visual arts, music, and imaginative writing at the middle and high school levels. The project grew out of a commitment to develop non-traditional models of assessment appropriate for students engaged in artistic processes. Its larger goal is to find means to enhance and document student learning in the arts and humanities. Supported by the Arts and Humanities Division of the Rockefeller Foundation, PROPEL was developed and field-tested during a five-year period from 1986-1991 by researchers at Harvard Project Zero and Educational Testing Service working in close collaboration with teachers and administrators in the Pittsburgh public school system.

For information on Arts PROPEL, contact:

Drew Gitomer
Mail Stop 18-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
609-734-1528

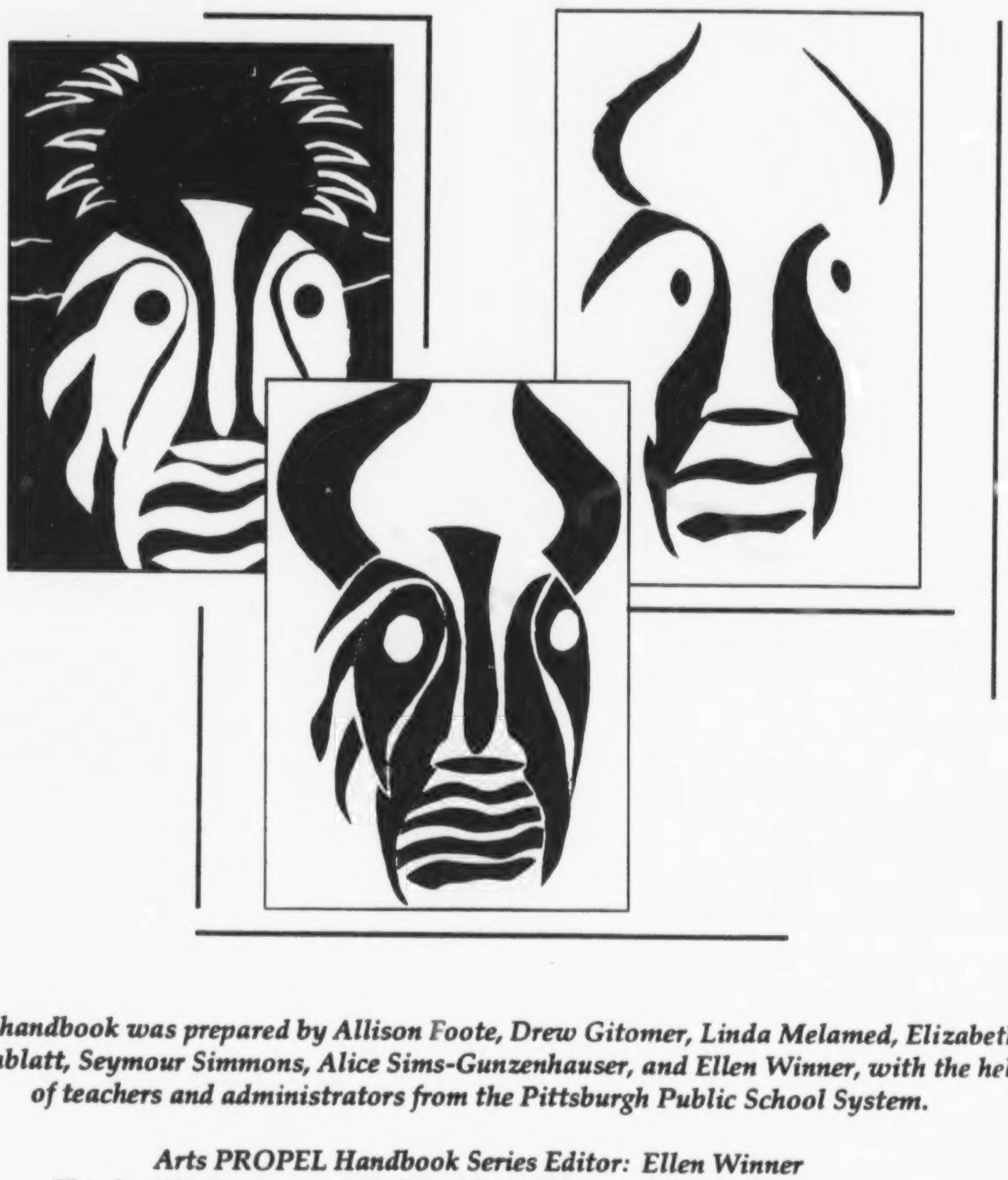
For publications, contact:

Marilyn Ispanky
Mail Stop 37-B
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
609-734-5073

These materials were reproduced with the permission of Educational Testing Service.

ARTS PROPEL:

A HANDBOOK FOR VISUAL ARTS



This handbook was prepared by Allison Foote, Drew Gitomer, Linda Melamed, Elizabeth Rosenblatt, Seymour Simmons, Alice Sims-Gunzenhauser, and Ellen Winner, with the help of teachers and administrators from the Pittsburgh Public School System.

Arts PROPEL Handbook Series Editor: Ellen Winner
This handbook was co-edited by Ellen Winner and Seymour Simmons.

A LOOK AT A MIDDLE SCHOOL PORTFOLIO

Janelle Hirschkopf was an 8th-grader in Pam Costanza's class at Rogers School for the Creative and Performing Arts in Pittsburgh. Pam Costanza teaches in what might be considered an ideal environment for developing a PROPEL approach to art education. She is surrounded by supportive colleagues and administrators who share her belief about the importance of art in education and who, in many cases, are similarly involved in PROPEL. The students in her classes are also unusual in that, by the time they enter the sixth grade, they have made a serious commitment to the study of art. They attend art classes during their entire three years at Rogers. Sixth graders take art for a 40 minute period two to four times per week. Seventh and eighth graders attend classes four days a week, and spend three consecutive periods, or two hours and fifteen minutes each day, in art. Seventh and eighth graders alternate regularly throughout the year between classes in two-dimensional media taught by Pam and classes in ceramics taught by an adjunct teacher.

The portfolio shown here begins with two of the many sketches Janelle did as weekly homework assignments. Figure 7.5 shows her first sketch, a drawing of a porcelain figurine. The figure is rich in detail and challenging in terms of its proportions and surface qualities. In later drawings, Janelle continued to focus on rendering details and surface qualities, but now concerned herself with more subtle issues of texture and tone as she began to "blow-up" small objects to several times their natural size (see Figure 7.6).

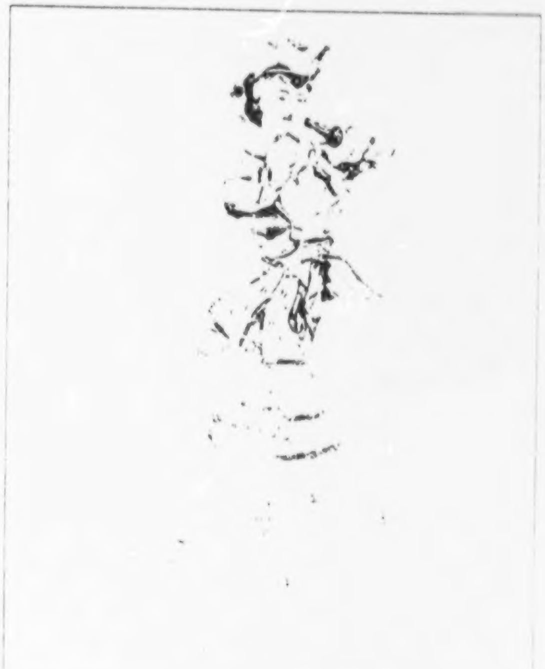


Figure 7.5

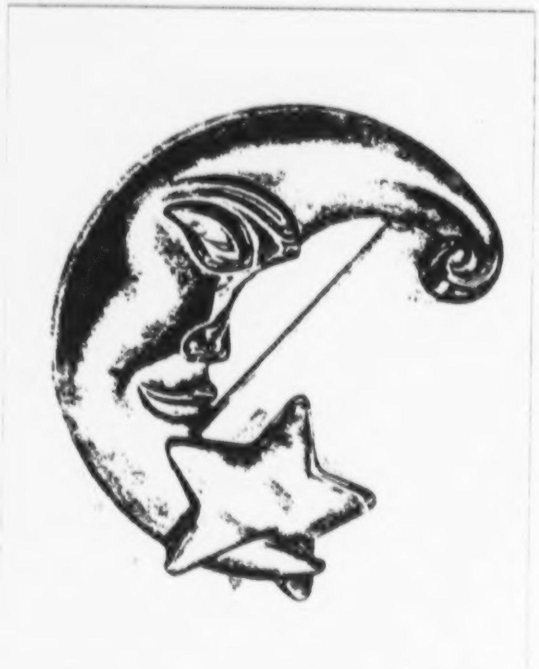


Figure 7.6

When asked in her end-of-term self evaluation which works she felt were particularly successful, Janelle said:

> *Pencil drawings are what I'm the most successful with. I've been using pencils basically all my life. I can really create some great stuff with it when I want to because I know what I'm doing when it comes to that kind of medium.*

She felt that her most frustrating experience was a watercolor assignment done around the same time as Figure 7.6. The frustration was due to feeling that she was unable to control the medium.

The first in-class two-dimensional domain project was a portrait unit. Building on portrait drawing experiences from previous years, this project was structured to help students learn to do portraits using a variety of styles and media. Students started with a series of three-minute line drawings, using felt tip-pens and drew the person across the table from them. Students made a blind contour drawing, a drawing using only circular lines, and a drawing using only straight lines made with the help of a ruler (see Figures 7.7, 7.8 and 7.9).

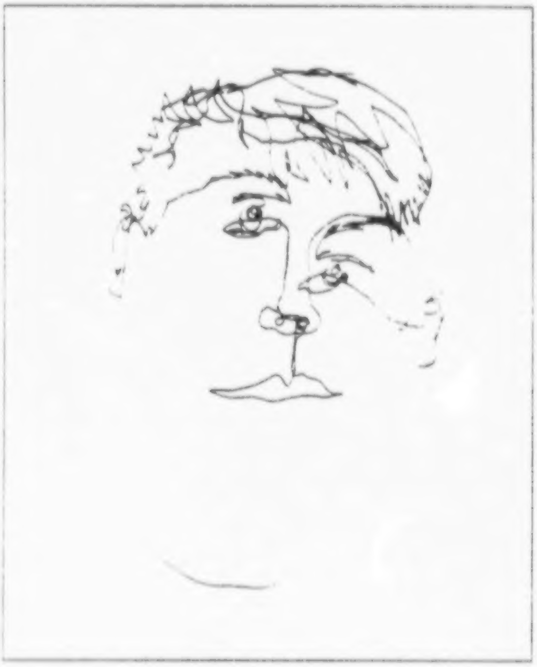


Figure 7.7

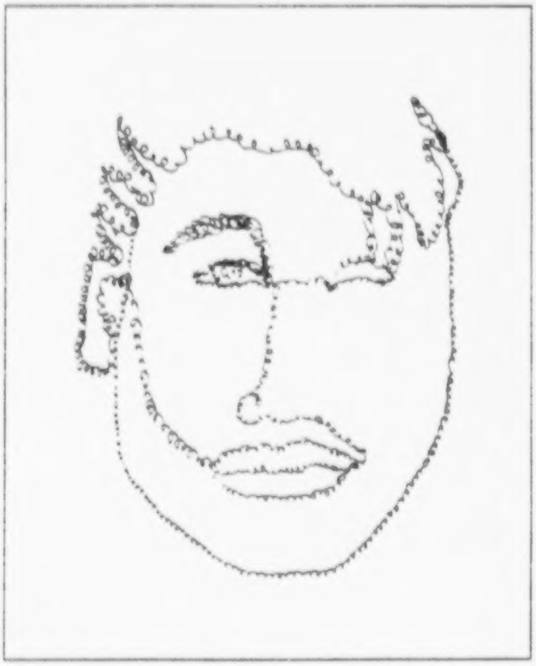


Figure 7.8

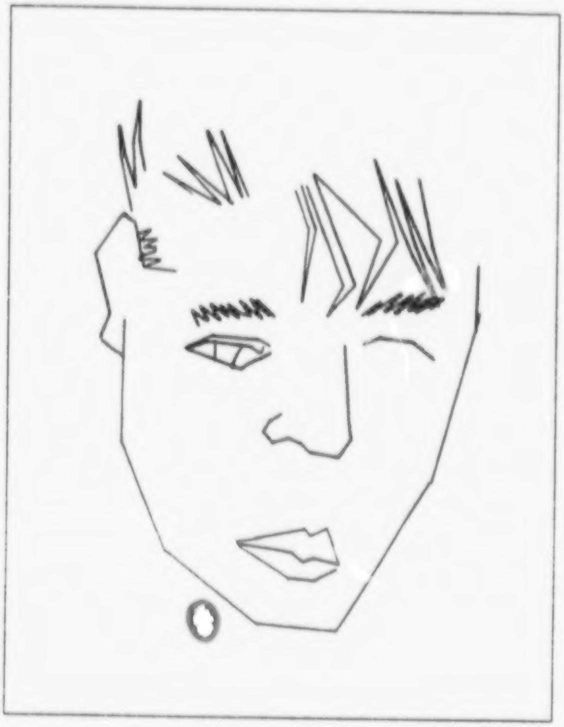


Figure 7.9

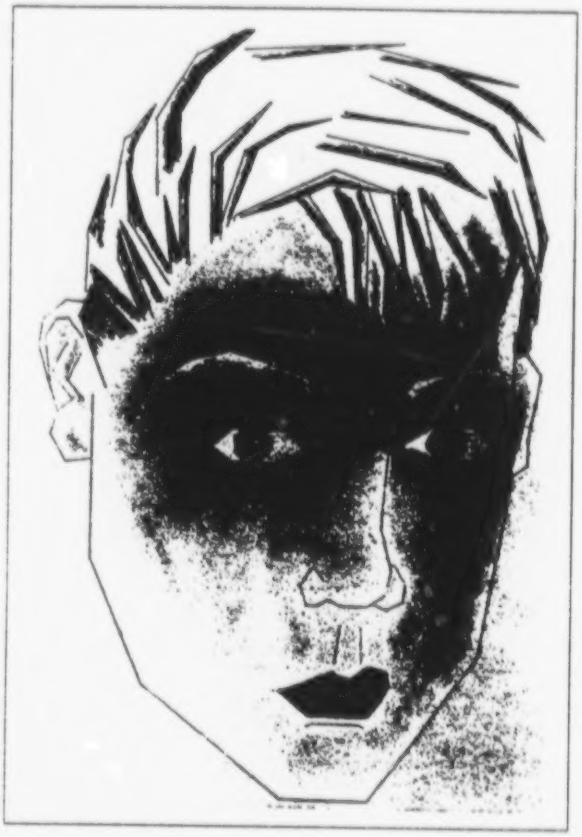


Figure 7.10

For each drawing, students made entries in their journals about how they felt about the drawing and the assignment. After all three drawings were done, they were put on the board to critique and discuss. A fourth drawing was then begun, initially using the ruler again to define the contours, then using oil pastels to paint in the face with "expressionist" tints that portrayed the person's personality. (Figure 7.10).

The completed oil pastels were put on the board for further discussion. At the end of the project, students looked at a series of Cubist and Expressionist portraits by Picasso, and were asked to discuss their portraits in comparison to those by Picasso.

The portrait project resumed, after several weeks spent on another assignment, with portrait drawings done in conte crayon and self-portraits done in pencil (see Figures 7.11 and 7.12). These were also critiqued in class. Before starting a final portrait drawing, students looked at and discussed portraits by artists with very different styles: Botticelli, Filippino Lippi, Vermeer, Rembrandt, Cassatt, Van Gogh, Modigliani, and Picasso. Students selected one or more artists as models, and created a portrait or self-portrait using the model as inspiration. Students could choose any of the media they had used during the project. As they began work, students recorded in their journals the objectives of the project, their chosen artist(s), subject, and media. These entries were accompanied by an explanation for the various choices.

BEST COPY AVAILABLE

Janelle wrote as follows:

> *I am going to do a self-portrait in colored pencil... because I want to learn how to use color to make portraits more interesting. The reason why I'm doing myself is because,if I want to express myself in any certain way by reflecting my personality, I know more about me, than I know about other people in this room.*
>
> *The style will resemble Botticelli's technique because I like how he makes his models pose. I had to change my medium to pastels. It is going to be more of a challenge but I guess it will make this project more exciting. I'm doing it from shoulder up so you can focus your attention mainly on the face, and since I'm starting a new medium it would hold me back to worry about the body on top of a new medium...In the background I'm going to put in deep blue sky with lots of clouds simply because I'm a daydreamer. I love to just sit around doing nothing but I am thinking and dreaming.*

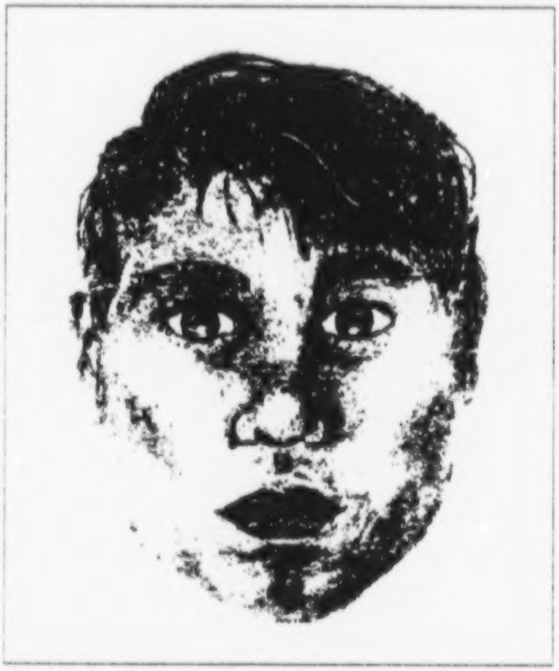


Figure 7.11

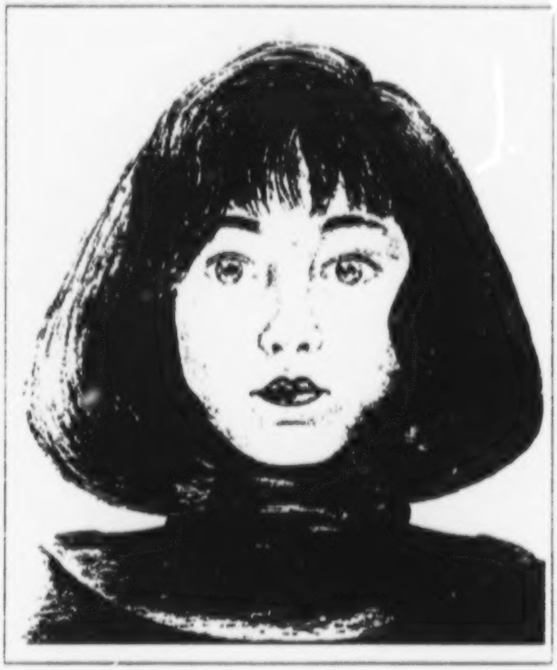


Figure 7.12

After doing the portrait, students critique their work in light of the assignment and their intentions (see Figure 7.13). These comments are summarized in the final portfolio review form.

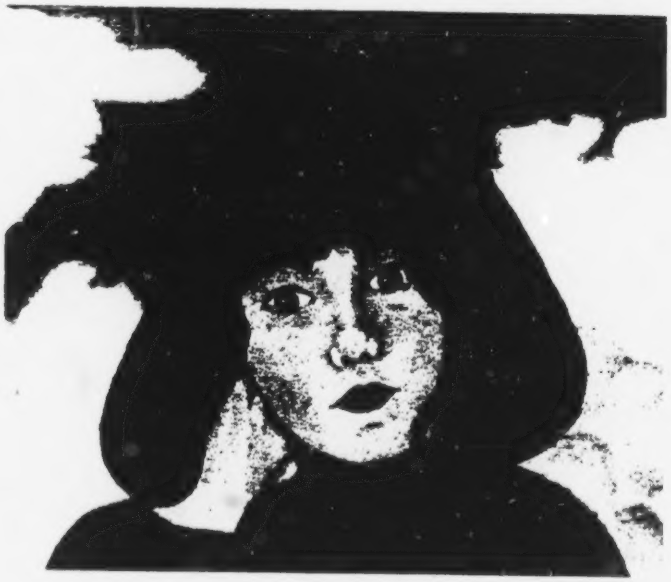


Figure 7.13

Although her teacher was pleased with the final self-portrait (above), Janelle was not. When asked, in her final portfolio review, to select a work she felt was not completely satisfying, she chose this one, saying:

> *I don't think it is expressive enough. The main reason why I picked a self-portrait is because I wanted it to reflect me. I thought I would be able to show myself through art but I didn't do too well.*

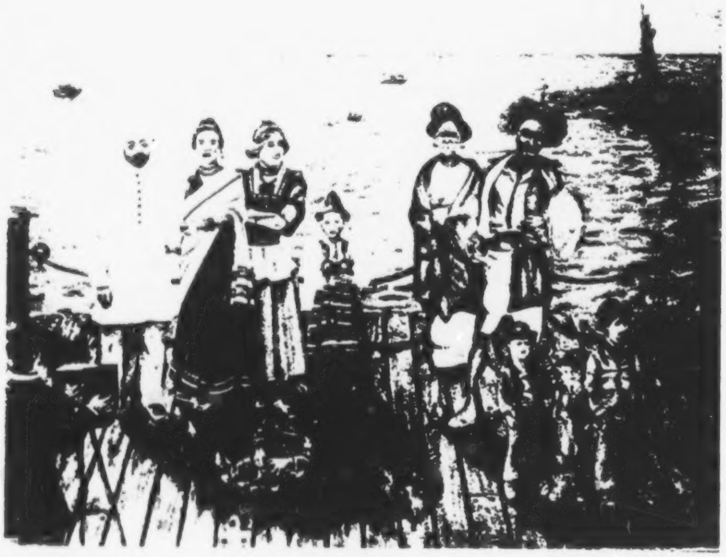


Figure 7.14

BEST COPY AVAILABLE

A series of diverse activities followed the portrait unit during the second half of the year. Among these were a watercolor unit, a Native American project, and a project done in conjunction with a social studies unit on "immigration and integration." For the last project, Janelle worked in collaboration with another student on a large pencil drawing showing immigrants wearing the costumes of many cultures arriving at the port in New York (Figure 7.14).

The final project for the year was determined by each student individually, the primary requirement being that it draw upon material and concepts taught during the year. For this assignment Janelle did another large drawing, this time, in pen and ink. The title of the drawing was "Family." (Figure 7.15)

Explaining her title, "Family," Janelle said:

> *Because that's what it is. I didn't want to say "the funeral" because then people won't be able to make up their own story to it. So with, say, just "family" you can come up with many different opinions and stories.*

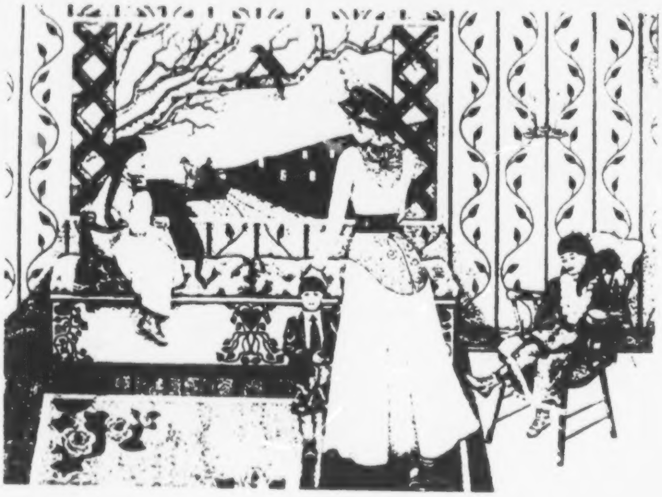


Figure 7.15

Commenting on her weaknesses, Janelle writes:

> *Everything blends in with other things around it. I'm taking too long. I can't get all of the shadows to make sense.*

About her strengths, she writes:

> *The people look like people except the girl in the chair has a beard. The stained glass windows' shades all look right. I managed to get patterns done easily, like on the window seat, and the wallpaper. Usually everything would be different, but these all are the same.*

JANELLE'S JOURNAL

In Janelle's journal are her drawings, along with her comments, and observations, reflecting her present and future concerns as an artist. She also included pictures and writings that inspired her, such as cut out images of faces from magazines, antique post-cards and photographs inherited from an elderly neighbor, and lengthy articles on portraiture and costumes apparently xeroxed from an encyclopedia. There are also extensive drawings and quotes taken from a book on anatomy. Although many of these images were eventually used for class projects, they were initially selected just because they intrigued Janelle.

Other drawings in the journal include sketches and cartoons which show a lighter and freer side of the student artist than one might expect looking at her class work (see Figure 7.16 (cat) and Figure 7.17 (cartoon). But also evident is a deep sensitivity to art and experience, captured in reflections on her work and in observations such as those cited in the box on the next page.

Pam collects the journals once each grading period and writes extensive comments, often in response to the student's observations.

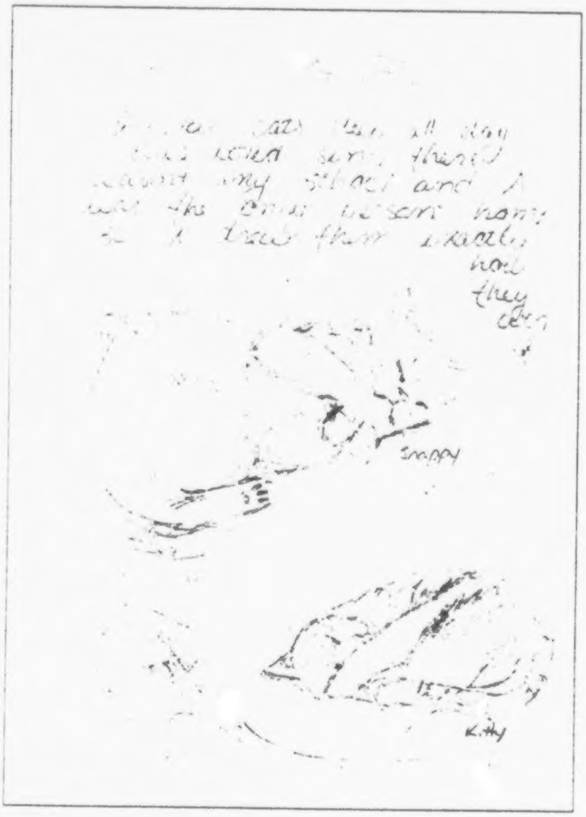

[illegible] cats [illegible] all day [illegible] my school and I [illegible] them exactly how they [illegible]

Figure 7.16

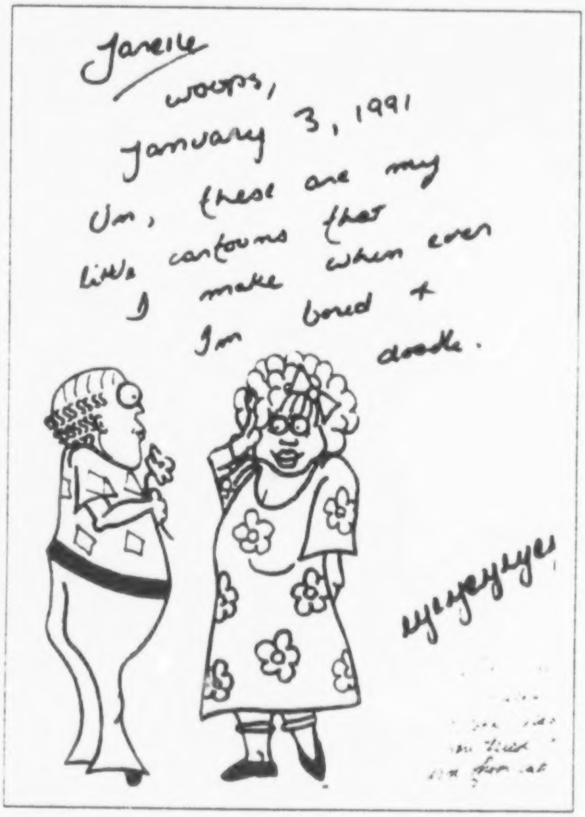

Janelle
woops,
January 3, 1991
Um, these are my little cartoons that I make when ever I'm bored + doodle.

Figure 7.17

BEST COPY AVAILABLE

A PAGE FROM JANELLE'S JOURNAL

October 23, 1990. Sometimes, I don't think I'll ever fully understand what Art is. There is so much of it and it comes in so many different forms and appearances. Lots of it is beautiful and lots of it is ugly. Many different feelings come with it, from graceful to clumsy, boring to shocking.

Art is everywhere and can be created from anything. Whenever I am riding through the city at night, I see all the buildings shining gloriously against the dark sky, the red tail lights glowing in front of me, and the passing shadows that are reflected from telephone poles, abandoned stores and other objects that line the city streets, I think to myself "that would be one of the coolist pictures in the world to be captured on canvas." Then there is the late afternoon scene. Maybe it just stopped raining and the appearance of everything has been grayed and misted. Bright warm mornings in the summer. Cold, black winter nights when the stars appear so sharp they cut through the sky.

Everything gives you different feelings and thoughts. That's why I like to draw people so much. Everyone has its own interesting personality and identity. Old people, young people, black, white, short, fat, tall and skinny. I hope that soon I will become a good enough artist to capture every intricate detail on my models. I hope to give them personality, feelings, moods, and appearances all with my pencil. But it will take a lot of hard practice.

In response, Pam wrote:

> *These are beautiful and sensitive observations. "What art is" is a question that changes constantly through the years. That's why it's a good idea to record each year what you think it is and how it changes from time to time."*

Assessment of Janelle's Portfolio

Pam provided the following mid-semester evaluation of Janelle's portfolio:

Production: When comparing these two sketches (referring to Figures 7.5 and 7.6), one can see how Janelle's drawing skills have improved. The sketch done in September was of a porcelain figurine. It is meticulous, detailed, and richly textured. The size of the sketch was the same as the object itself. The January sketch is a more refined rendering of a pin, an object much smaller than the drawing. Janelle has captured the smooth "buffed" metal and "highlighted" the reflective part of the eyes, nose, lips, and end of the moon which were highly polished, demonstrating a greater sophistication.

Reflection: In Janelle's final portrait (Figure 7.13), she wanted it to express her personality. Although she did a beautiful rendering, she was not pleased; it did not accomplish her goals. There are many changes she would make that would express her personality more accurately.

The following are some of Janelle's comments about the changes she would make which were excerpted from a Portfolio interview with Pam:

> *"I didn't put enough feelings into it and I think I need to change the background because I'd put more feelings into it . . . I could change the pose a little bit . . . the skin's too pale. It just looks possessed. Everything stands out but the skin . . . I think I'd want to put more of my body in it so I can have it show more of me, other than just my head. Maybe a different face expression . . . I'd probably smile or something.*
>
> *The reason why I put the background in there is because I do daydream a lot, but I don't do that all the time. I have something more wild and something more alive than just sitting there. I wouldn't put shapes because everybody puts shapes to express themselves. I'd have some kind of weird scenery . . . and I'd want people in the background.*
>
> *I wouldn't make it so still. It just sits there and looks at you. It doesn't have anything to it. I'd make it more alive, put more colors into it. Make it more colorful and bright . . . I'd have my hair flying around everywhere."*

Perception: I think the body of Janelle's work thus far has shown a keen sense of perception of her environment, from the subjects she chose for weekly sketches, to her self-portrait.

Multiple Challenges

A Series of Questions Illustrating the 1992 National Assessment of Educational Progress

This packet illustrates the diversity of the questions on NAEP's 1992 mathematics, reading, and writing assessments. Over more than 20 years, NAEP has consistently used constructed-response questions to augment the multiple-choice format — including hands-on tasks in science and mathematics; literacy tasks involving newspapers, charts and tables, bus schedules, and pay stubs; interviews in reading and lengthy writing tasks.

For more information on NAEP, write to:

PO Box 6710
Educational Testing Service
Princeton, New Jersey 08541-6710

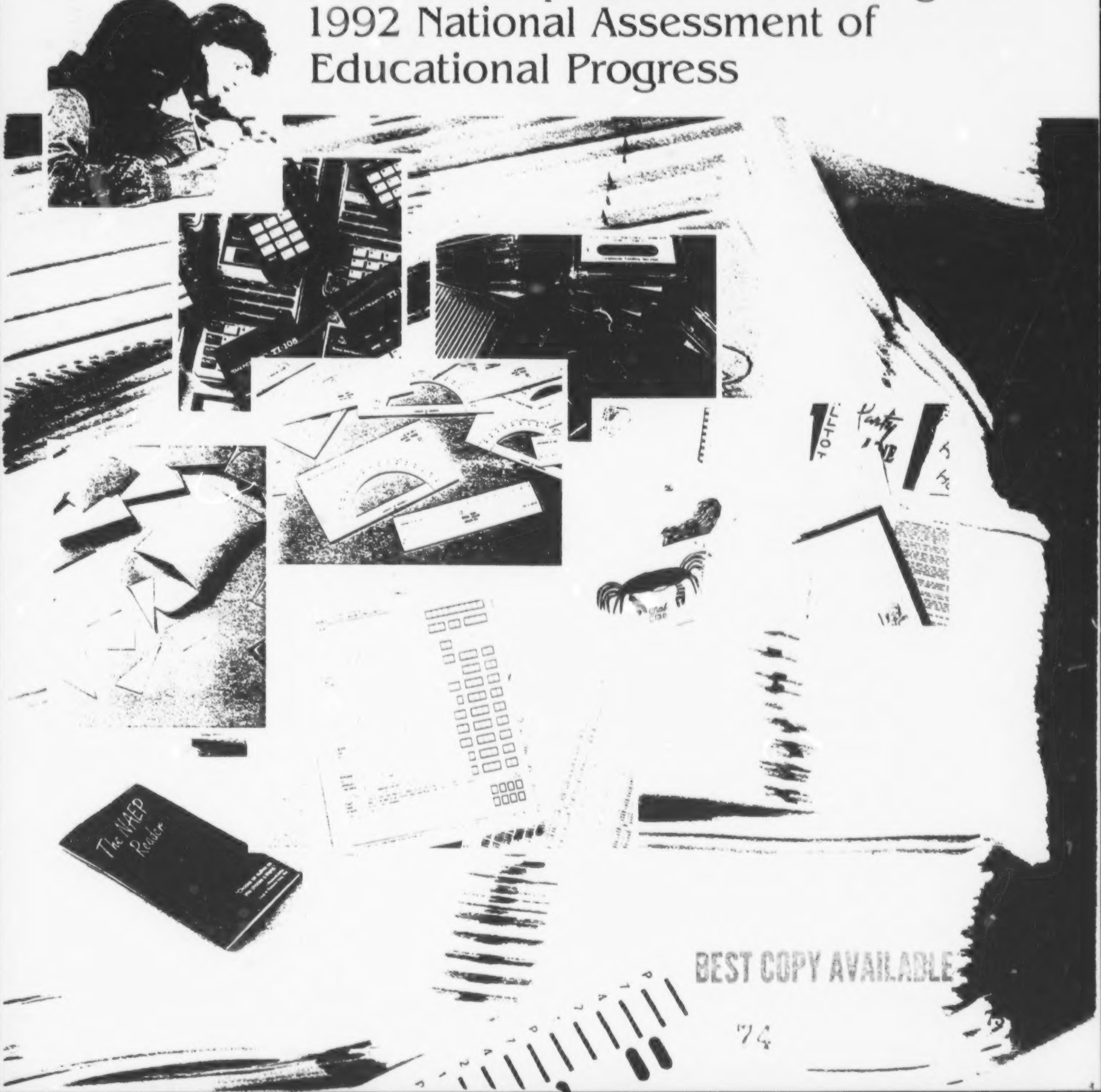
or to:

Education Information Branch
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Ave., NW
Washington, DC 20208-5641

THE NATION'S
REPORT
CARD
naep

Multiple Challenges

A series of questions illustrating the 1992 National Assessment of Educational Progress



BEST COPY AVAILABLE

74

Children require not only that secure learning, but a good measure of new experiences, from flames to flowers, all actively present and looked at questionably in the schoolroom. The stuff has to be brought into schools by teachers and their supporters.

Philip Morrison

(discussing NAEP science assessment results, September 1988)

BEST COPY AVAILABLE

THE NATION'S
REPORT
CARD
naep

Multiple Challenges

This packet illustrates the diversity of the questions on NAEP's 1992 mathematics, reading, and writing assessments. We hope that teachers will find it informative as they consider performance assessment and develop their own classroom tests. As with the entire NAEP, the sample questions presented herein were developed with the thoughtful assistance of teaching professionals.

NAEP is designed to present students with an array of innovative tasks that make learning come to life. The range of challenges on the assessment parallels the complexity of the future. Students will be **contemplating** and **adapting** to new environmental conditions and mastering mathematical and technological concepts and skills. They'll be **reading**, **discussing**, and **elaborating** on people and events in literature. They'll be **organizing** and **synthesizing** volumes of information and using a variety of informative and persuasive techniques to communicate.

During its 22-year history, NAEP has consistently used constructed-response questions to augment the multiple-choice format — including hands-on tasks in science and mathematics; literacy tasks involving newspapers, charts and tables, bus schedules, and pay stubs; interviews in reading, and lengthy writing tasks. The students invited to participate in the 1992 assessment will be asked to address a variety of thought-provoking questions — giving policymakers and the general public useful information on what American schoolchildren know and can do.

Since the framework for NAEP assessments is periodically updated to reflect current thinking in each field, NAEP's objectives and reports are a valuable resource for teachers developing instructional activities that are both innovative and intellectually challenging. The 1992 NAEP (in mathematics, reading, and writing) includes diverse performance tasks designed to examine student achievement in grades four, eight, and twelve.

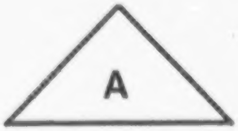
Mathematics

The National Council of Teachers of Mathematics has set standards for mathematics education that suggest students should be able to use mathematics as a way to solve practical problems, to communicate mathematical ideas to others, and to reason properly. Using technology in the classroom can accelerate the pace of student learning and help make school mathematics more like the mathematics people use in their everyday lives and on the job.

The NAEP mathematics assessment — which requires students to use calculators, rulers, and protractors — contains questions that direct students to sketch, measure, and manipulate geometric figures; to represent algebraic equations graphically; and to give brief written explanations to support solutions to problems. The framework for the assessment is organized according to mathematical abilities and content areas. The mathematical abilities assessed are **conceptual understanding, procedural knowledge, and problem solving**. The content areas assessed are **Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions**.

The constructed-response questions provide an extended view of students' mathematical abilities that cannot be measured using multiple-choice questions. These include the ability to articulate mathematical ideas, make estimates, develop informal proofs, draw figures, and generalize relationships.

The following are examples of geometry tasks: number 1 for fourth grade; numbers 2, 3, and 4 for eighth and twelfth grade; and number 5 for twelfth grade.



(The students would be given the cut-out cardboard triangle drawn above and a straightedge.)

1. **Piece A represents what geometric figure (shape)?**

Isosceles triangle, isosceles right triangle, triangle, right triangle.

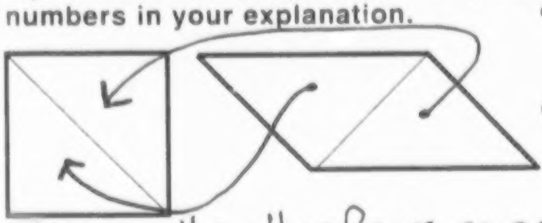
2. **Name three geometric properties of piece A.**

It has 3 sides (3 angles).

It has 2 equal sides (2 equal angles).

It has 1 Right angle.

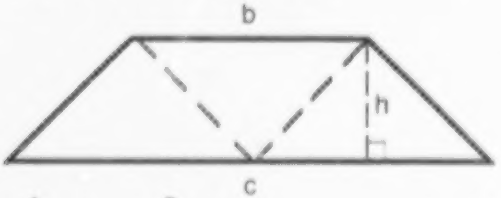
3. **Use piece A to show that the two figures below have equal areas. You may use drawings, words, and numbers in your explanation.**

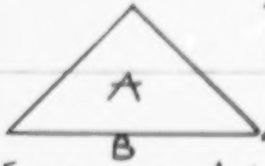


Pieces from one figure fit exactly over pieces on the other figure, so areas are equal.

---

5. **Use piece A to prove that the area of the figure below can be found by using the formula $A = \frac{1}{2}h(b+c)$. You may use drawings, words, and numbers in your explanation.**

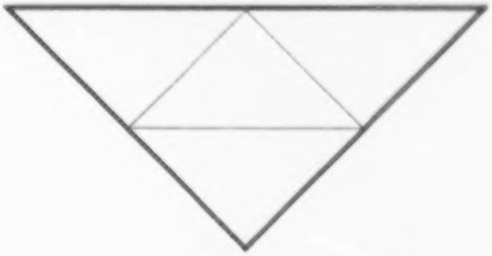




Three piece A's fit exactly into the figure.

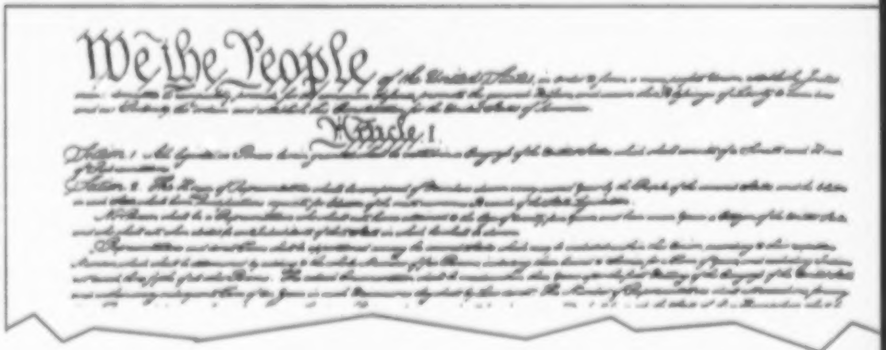
Each piece A has area $\frac{1}{2}bh$. The figure's area is $\frac{1}{2}bh+\frac{1}{2}bh+\frac{1}{2}bh = \frac{1}{2}h[b+b+b] = \frac{1}{2}h[b+c]$, since $b+b=c$.

4. **If the area of piece A is 3, what is the area of the figure below? Explain how you found your answer. You may use drawings, words, and numbers in your explanation.**



The area is 12 since the figure is made up of four identical triangles like A and 4×3=12.

Reading


We the People

Article. I

Recognizing that readers involve themselves differently with different material, depending on the kind of text and the purpose for reading, the 1992 NAEP reading assessment uses a variety of passages to assess specific kinds of reading. NAEP examines performance in three basic reading situations — reading for literary experience, reading to obtain information, and reading to perform a task. For example, we read novels for literary experience, newspapers for information, and instructional manuals to accomplish tasks. As readers' purposes change, so do their approaches to the material, even when using the same text.

The assessment allows students either 25 or 50 minutes to read and respond to questions from reading material drawn from published sources. Within each of the following situations, certain reading abilities are assessed – every reader needs to draw on them to read effectively.

- **Building an understanding** of a passage includes forming an initial, global understanding of it. To assess this understanding, NAEP asks constructed-response questions like "What does this passage say to you?" or "What does the author think about the topic discussed in the paragraph?" Usually, the first question taps the reader's initial impression of the text; follow-up questions deal with a more in-depth understanding.

- **Developing an interpretation** goes beyond the initial impression to seek more complete understanding. Assessment questions that measure this ability include: "How does this character change from the beginning to the end of the story?" or "What caused the little girl to get angry?"

- **Personal reflection and response** requires the student to relate the text to personal knowledge or experiences. Questions that require personal reflection and response include "How is this story like or different from your experiences?" or "What additional information would you like to know about this topic?"

- **Demonstrating a critical stance** requires the student to consider the text objectively. It involves evaluating, comparing and contrasting, applying knowledge, and understanding text features such as the author's use of irony and the organization of ideas. Some critical stance questions require readers to make connections across parts of a text or between texts. Questions that require a critical stance include "What could be added to improve the author's argument?" or "Poem A and Poem B have similar themes, but how are they different?"

In 1992, NAEP is conducting a special study of students' oral reading skills called the **Integrated Reading Performance Record (IRPR)** – for a description, see *Special Studies in Reading & Writing* in this booklet.

The following are examples of an eighth-grade reading task to assess informational experience and a fourth-grade reading task to assess literary experience.

The Constitution of the United States

There is a building in Washington, D.C., called the National Archives, where our most valuable historic documents and materials are stored. Journals and newspapers, personal letters, photos that date back to the time when photography was invented—these are some of the Archives' important treasures. But there are two documents in the Archives that hold a place of special honor. Housed in a display area that is itself something of a marvel, these two documents attract thousands of visitors every year. They are our Declaration of Independence and the Constitution of the United States.

The Declaration of Independence, faded and barely readable, and the Constitution, in excellent condition after almost two hundred years, attract visitors not because they are old and beautiful to look at, but because they represent the basic ideas that Americans live by. When visitors have the opportunity to view the actual documents, many feel in awe of what the documents stand for. They were written at a time when Americans were struggling to create a new and lasting government. Men and women—including John and Abigail Adams, Thomas Jefferson, Alexander Hamilton, Ben Franklin, and James and Dolley Madison—were studying, debating, and presenting arguments for different kinds of government. The basic ideas that affect how we live today were set down from 1776, when the Declaration of Independence was issued, to 1788 when the Constitution was ratified.

The Stage Is Set For the Convention

The period following America's War of Independence was a difficult time for our country. During the war, the Continental Congress, which was the assembly of people

On the basis of what you know from this article and your personal knowledge, explain whether the Articles of Confederation helped or hindered progress toward a new form of government.

I think the articles of confederation helped the progress toward a new form of government because if they wouldn't have failed at first we might not have ever thought of a new way to form the government the way we did. The war a disaster could remain because of the high taxes and the paper currency and they couldn't find a way to raise money so they gave up and started over. I think this helped because now they had some idea of how they could and couldn't run their government as "the saying goes," If at first you don't succeed, try, try again", and that's what was done back then.

I'll Fix Anthony

by Judith Viorst
Illustrated by Arnold Lobel

My brother Anthony
can read books now,

but he won't read
any books to me.

He plays checkers
with Bruce
from his school.
But when
I want to play
he says Go away or I'll clobber you.

I let him wear
my Snoopy sweat shirt,
but he never lets me
borrow his sword.

Mother says deep down
in his heart Anthony loves me.

Anthony says deep down
in his heart he thinks I stink.

When I'm six a dog
will follow me home,

and she'll beg for me and roll over
and lick my face.

If Anthony tries to pet her,
she'll give him a bite.

When I'm six Anthony will have
the German measles,

and my father will take me
to a baseball game.

Then Anthony will have
the mumps.

and my mother will take me
to the flower show

Think about the plan to "fix" Anthony. Write a paragraph explaining whether or not this plan is a good idea. Give examples from the poem.

This plan is not a good one because he doesn't realize that when he is six Anthony will still be older and bigger. For example what is the point of putting his sneakers on the shelf because by the time he is six Anthony will be able to reach the shelf himself. By that time Anthony will be to mature to play his brothers little games. Beside which when he is six he'll probably forget this whole thing ever happened. Like in the poem it shows him in a car, he doesn't know six-years-olds aren't aloud to drive a car.

Writing

The writing assessment requires each student to write either two 25-minute essays or one 50-minute essay. Topics range from informative to persuasive to narrative and ask students to generate ideas, synthesize information, and organize their thoughts. Each task is accompanied by a "planning page" for students to use in jotting down ideas, diagramming, or outlining their thoughts. NAEP then evaluates the quality and fluency of these responses and monitors the trends in students' writing performance across the years. Students' responses are evaluated on the basis of their success in accomplishing the specific purpose of each writing task, as measured by an enhanced application of "primary trait" scoring. Based on a six-point scale, the evaluation criteria measure students' success in selecting, organizing, and presenting relevant information as well as their use of effective organizational and development strategies (e.g., compare, contrast, and anecdote).

In 1992, NAEP is conducting a special study of classroom-based writing called **The Nation's Writing Portfolio** – for a description, see *Special Studies in Reading & Writing* in this booklet.

The following are examples of eighth- and twelfth-grade writing tasks, respectively.

Think about a favorite story you have read or heard, or one that you have seen in the movies or on television. Write a paper, telling what the story is about and why you like it. Help other people to understand why you think that it is a good story. Use examples from the story, such as details about characters, places, events, or ideas.

Write your paper on the lined pages

For 21 Jump Street I saw a show I really liked. This cop was supposed to go undercover to a regular school. People had been selling drugs there, and it was her job to find out who it was. She dressed like a regular kid, and acted like a regular student and no one expected a thing. She met a few girls and they offered her some drugs. She didn't know who was pushing the drugs, but she intended to find out. The school was debating if they should have a drug test. This way they would know who had problems, and they could get them help. When the cop got her test back it showed positive. She was shocked. She had never done drugs before she went to the [illegible] where she worked. They knew she was innocent but they had to take her badge because it was the law. She still stayed on the case even though she wasn't being paid for it. She went to the hospital to take another test so she could clear her name. They had to cut some of her hair which would show them if she had taken any drugs. The tests came back negative and she was cleared. She got her badge back. She was still determined to find the pusher and she did.

I thought it was a great show. It showed kids that drugs are bad for you, and there really are people out there willing to help you. Although most times tests are right, sometimes they can be wrong. You should always stand up to what you believe in and if you didn't do something that you were accused of, you should stand up to it. The story also showed that cops aren't bad people. They are just like you and me.

Your school has decided to employ students to fill several new part-time jobs at school. You want to gain more work experience and the pay is attractive, so you have decided to apply for one of these jobs. To be considered, you need to submit a letter of application, identifying the job you want and summarizing your qualifications for it. Many students will apply for each job, so you need to convince the employment director that you are the best candidate.

Many positions are available. For example, the school will hire students to paint classrooms, do library work, fix up recreational areas and equipment, help maintain a computer system, make props and costumes for special events, and help teachers prepare materials for their classes. Students can also suggest jobs that they think are needed at school. Students can perform these jobs either during the school year or during the summer.

According to the employment director, your application letter should include the following elements:

- a short description of the job you would like to have at school;
- a summary of the skills that you think are needed to do this job well, and
- personal information on previous work experience, job-related interests, future employment or educational plans, and other relevant details.

I believe I would be very good for this job because I have had a lot of experience in working with computers. I have taken two full years of typing with a typing rate of 45 words per minute. Currently I am taking a computer course and learning many different ways to use a computer. Also my family has two computers at our house. One is for the family business and the other is for us to work with. Over the last two summers I have been taking programs working with computers and programming them. I also have a job working with a computer at an office firm. I have had other jobs working with computers. I believe I would be very good for this job. This is why I would like to work with the computer system.

BEST COPY AVAILABLE

Special Studies in Reading & Writing

As teachers devote more instruction time to presenting students with multiple challenges, NAEP continues to devise alternative methods of assessing how well students are meeting these challenges.

The 1992 NAEP assessments provide two special components that will extend and enhance the information from the main reading and writing assessments.

The Integrated Reading Performance Record (IRPR), a special study to augment the 1992 paper-and-pencil reading assessment, will yield detailed information on students' oral reading fluency and comprehension.

A representative subsample of fourth graders participating in the main NAEP reading assessment will be chosen to take part in audiotaped interviews conducted by trained administrators.

Students are asked to bring to the interview their current reading textbook, three samples of work completed for reading class, and a favorite book. They discuss with the interviewer their independent reading experiences and samples of their classroom work. Students also reread one of the passages they responded to in the written portion of the assessment and orally answer questions about it. While the student reads aloud, the administrator notes miscues, reading time, and the student's phrasing of text.

The IRPR, together with the paper-and-pencil assessment, provides a more in depth view of reading ability by allowing students several opportunities to demonstrate achievement.

The Nation's Writing Portfolio is designed to collect students' best papers. This study of classroom–based writing also describes the assignments typically given to students and the broad range of procedures and strategies students use to complete the tasks.

Through the portfolio assessment, students can demonstrate their skill at writing when they have the opportunity to edit and revise their work. NAEP also can evaluate the relationship between writing produced in the classroom and that produced under timed conditions.

Teachers, working with fourth and eighth graders who participate in the main NAEP writing assessment, will review and select three papers that best illustrate each student's achievement as a writer.

The papers will represent a range of types of writing tasks (stories, reports, essays, and persuasive pieces) and the use of writing process strategies (successive drafts, use of reference sources, and peer review).

Along with their papers, students will also attach a brief explanation of their choice of papers to be submitted. Teachers will fill out a short questionnaire describing the assignments and the instruction that led to each student's writing.

Analyzing the Portfolios
Each portfolio will undergo a three part analysis. First, a trained reader completes a descriptive coding sheet noting the type and form of writing, the audience, the number of words, evidence of revising, and other information. Second, the reader evaluates pieces classified as narrative, informative, or persuasive, according to criteria established by teachers. Papers are rated from 1 to 6. Then, using the teacher questionnaire and student letter, the reader synthesizes the information about the assignments and contexts that produced the pieces of writing.

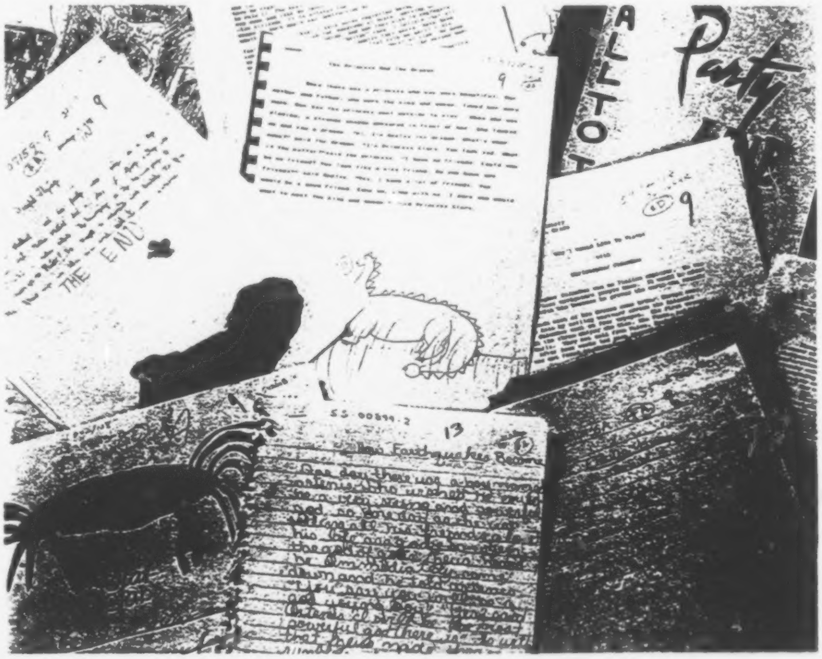
For narrative writing, a paper rated as 1 contains a list of sentences minimally related – while a rating of 2 is given to brief papers with a few details about settings, characters, or events. A 3 rating is given to a paper that describes a series of events but lacks cohesion due to

problems with syntax, sequencing, missing events, or an undeveloped ending. While a paper rated as 4 describes a sequence of episodes, including details about most story elements, it nonetheless is confusing or incomplete. A rating of 5 is given to a paper that, while describing a sequence of episodes in which almost all story elements are clearly developed, may have one or two problems or include too much detail. A rating of 6 represents a paper with a well-developed description of episodes, elaborated resolution of goals or problems, and cohesive presentation of events.

For informative writing, a paper rated as 1 lists pieces of information or ideas all on the same topic, but does not relate them, while a 2 rating represents a paper that relates a broader range of information without clearly establishing ideas, explanations, and details. A paper rated as 3 includes a broad range of ideas with some established relationships, while a 4 rating signals clearly related information and use of rhetorical devices. A rating of 5 is given to papers that present well-developed information and relationships with explanations and supporting details, while a paper rated as 6 has an overt organizational structure, a coherent sense of purpose and audience, and is free from grammatical problems.

For persuasive writing, a paper rated as 1 states an opinion but does not support it, while a 2 rating is given to papers that support an opinion but fail to explain the reasons. A paper rated as 3 attempts to develop reasons for the opinion with explanations that are not developed or elaborated, while a paper rated as 4 supports the reasons with explanations developed through the use of rhetorical devices. A 5 rating is given to a paper that contains an opposing point of view and an attempt to discuss and/or refute it, while a paper rated as 6 summarizes opposite points of view and clearly and explicitly refutes them.

Further information about the scoring guides for the National Assessment can be obtained from the project staff.



84

BEST COPY AVAILABLE

Learning by Doing

A Manual for Teaching and Assessing Higher-Order Thinking in Science and Mathematics

Intended for use by science and mathematics coordinators and teachers, this manual presents 11 tasks field-tested by NAEP during a pilot study in 1986. The tasks were selected to show a range of possibilities for both classroom and assessment use. Many of the ideas underlying the hands-on tasks can be adapted to a variety of different mathematics and science concepts. Each task is identified by the thinking skills necessary for successful student performance and the administrative mode used by NAEP. The presentation for each task includes a brief description of the activity, the student response sheet, a list of the equipment used, and one or more exemplary student responses.

Learning by Doing is adapted from *A Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics, Final Report.* This two-volume 537-page report, which describes NAEP's project in detail and presents all 30 tasks included in the pilot study — six group activities, 20 station activities, and four complete experiments — is available for $35 plus shipping and handling from:

NAEP
CN6710
Educational Testing Service
Princeton, NJ 08541-6710

Learning by Doing

A Manual for Teaching and Assessing *Higher-Order Thinking* in Science and Mathematics

THE NATION'S
REPORT
CARD
naep

BEST COPY AVAILABLE

Learning by Doing

A Manual for Teaching and Assessing *Higher-Order Thinking* in Science and Mathematics

May 1987
Report No: 17-HOS-80

From NAEP's Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics, supported by the National Science Foundation through a grant to the Center for Statistics, Office for Educational Research and Improvement, U.S. Department of Education.

The National Assessment of Educational Progress, *The Nation's Report Card*, is funded by the Office for Educational Research and Improvement under a grant to Educational Testing Service. National Assessment is an educational research project mandated by Congress to collect data over time on the performance of young Americans in various learning areas. It makes available information on assessment procedures to state and local education agencies.

The work on which this publication is based was supported by the National Science Foundation through the Office for Educational Research and Improvement under Grant No. NIE-G-84-2006-P4. It does not necessarily reflect the views of either of these agencies.

The ideas for the majority of the exercises for the project were taken from questions constructed for the national monitoring of science performance carried out by the Assessment of Performance Unit (A.P.U.) in the United Kingdom. We acknowledge the cooperation of the United Kingdom Department of Education and Science and of the unit in the Centre for Educational Studies in King's College, London, in making these questions available. However, the questions have been substantially changed to function within NAEP's very different framework. The U.K. A.P.U. is not responsible for the use NAEP has made of its ideas.

Library of Congress Catalog Card Number: 86-63907

ISBN 0-88685-059-2

Educational Testing Service is an equal opportunity/affirmative action employer.

Educational Testing Service, *ETS*, and are registered trademarks of Educational Testing Service.

Table of Contents

Introduction 5

Classifying
Task: Vertebrae 8

Observing and Making Inferences
Task: Wig-Wag.............. 10
Task: Whirlybird 12
Task: Conductivity 14

Formulating Hypotheses
Task: Water on Brick........ 16
Task: Double Staircase...... 18

Interpreting Data
Task: Triathlon............. 20

Designing an Experiment
Task: Heart Rate and Exercise 22

Conducting a Complete Experiment
Task: Sugar Cubes 24
Task: Survival 26
Task: Density 28

Thoughts for the Future 31

Acknowledgments 32

Introduction

Learning by Doing: A Manual for Teaching and Assessing *Higher-Order Thinking* in Mathematics and Science

This manual is designed for use by science and mathematics coordinators and teachers.

Why Hands-on Assessment?

Improving ways to teach and measure higher-order thinking skills has become a national priority, primarily because *A Nation at Risk* and other prestigious reports have identified a crucial need for more sophisticated skills among our nation's students. For example, *Educating Americans for the 21st Century*, the report of the National Science Board's Commission on Precollege Education in Mathematics, Science and Technology, stated, "We must return to the basics, but the basics of the 21st century are not only reading, writing, and arithmetic. They include communication and higher problem-solving skills, and scientific and technological literacy—the *thinking* tools that allow us to understand the technological world around us. These new basics are needed by *all* students. . . ."

The emergence of new jobs requiring technological skills and expertise, concern about the world environment, and the need in our daily lives to make important decisions based on new medical and scientific discoveries have also served to heighten interest in science and mathematics education. Although all schools require some mathematics, student participation in science courses is not widespread in American schools. This is particularly true in elementary schools, where, according to the Association for Supervision and Curriculum Development, a typical fourth-grade curriculum allots only 28 minutes per day to science. Preliminary data from NAEP's 1986 science assessment show nearly one-fourth of the third graders reported that they rarely or never had science class. Even in the higher grades, students did not report taking a variety of science courses. While many eleventh graders reported having taken biology, less than 40 percent had taken chemistry and only about 10 percent had taken physics.

This relatively low participation in science courses suggests that many students may have limited experience with laboratory or hands-on applications of scientific and mathematical concepts. Students should have both the concepts and process skills necessary to organize and carry out projects in an increasingly complex world. Hands-on instructional activities give them the opportunity to use knowledge and skills to solve problems and find out how and why things happen. Further, it is critical that assessment procedures be consistent with the best of these instructional practices. In *First Lessons*, U.S. Secretary of Education William J. Bennett writes:

> "The problem of assessment also constrains the spread of 'hands-on' science. It is relatively easy to test children's knowledge when they have been asked to memorize lists of data for a test. It is much harder to design tests that measure learning derived from direct experience; some school systems provide checklists of students' ability to perform experimental tasks. The challenge before science educators is to develop better means of measuring both factual knowledge and the kinds of understanding students acquire through activities. When that task is accomplished, a major roadblock to science achievement will have been removed."

NAEP, the Nation's Report Card, has developed and pilot-tested a variety of hands-on science and mathematics tasks. These tasks were developed as prototypes for use in future national assessments, but the concepts measured and the innovative approaches used are equally suitable for classroom learning. This manual is designed to share these techniques.

What are the tasks like?

The tasks presented in the following pages require students to think independently about a variety of relationships in mathematics and science. At the first level of the hierarchy, students are asked to **classify** and **sort** by identifying common characteristics of plants and animals. At the next level, students are given materials, equipment, and/or apparatus that exemplify particular mathematical or scientific phenomena or relationships and are asked to **observe, infer**, and **formulate hypotheses.** Another set of tasks is designed to measure students' ability to **detect patterns** in data and **interpret** the results. At the most complex level, students are asked to **design and conduct complete experiments.**

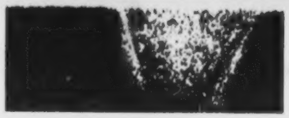
How were the tasks developed?

To develop hands-on activities asking students to solve problems, conduct investigations, and respond to questions using materials and equipment, NAEP invited the views of many science and mathematics educators. NAEP also worked closely with members of the United Kingdom's Assessment of Performance Unit and their science-monitoring staff at Kings College, London University. Many of the tasks were adapted from those used successfully in England, Wales, and Northern Ireland.

How were the tasks administered?

Because a major goal of this pilot project was to judge the feasibility of more innovative and complex assessment procedures, NAEP developed prototypes of different administration formats, including paper/pencil tasks, demonstrations, computer-administered tasks, hands-on tasks, and various combinations of these formats. These were grouped into three major administration modes.

1. ***Group activities*** were administered to intact classes. These tasks asked for open-ended paper/pencil responses to problems posed in various ways. One task included a demonstration of an experiment by the exercise administrator. The remaining tasks were based on various types of written or tabular information.
2. ***Station activities*** were hands-on tasks that required students to use equipment or materials to investigate relationships and then answer open-ended questions based on their findings. These activities were divided into two sets of six tasks for each grade level. Groups of six students were given the tasks, with students rotating from activity to activity every eight minutes. One task in each of the sets was administered by computer. Students received directions and recorded their answers by using the computer.
3. ***Complete experiments*** were administered to individual students. The administrator posed the questions, explained the equipment, and used a checklist to record how students used the equipment to conduct their experiments. After students had completed their investigations, they discussed their findings with the administrator.



Who participated in the pilot testing?

Twelve school districts across the four regions of the country participated in the pilot project. Within each region NAEP selected schools in middle-income urban, disadvantaged-urban, and small-city areas. Twenty-two trained administrators assigned in teams of three conducted the pilot study during April 1986. About 1,000 third-, seventh-, and eleventh-grade students were assessed, with approximately 100-300 responses obtained for each task.

What did the results show?

NAEP collected the pilot data primarily to assess the quality and grade-level appropriateness of the tasks rather than levels of student performance. From this perspective, the findings served their purpose. They indicated that students responded to the tasks, and in some cases, did quite well. Also, the results conformed to expectations about basic developmental trends in thinking skills. For example, improved levels of performance were observed across all three grade levels, and—given the grade-appropriateness of the tasks—students had less difficulty with the sorting and classifying tasks than with determining relationships and conducting reliable experiments.

However, staff and consultants wanted to know much more. The promise of new information obtainable from a hands-on national assessment was perhaps the source of most enthusiasm. Questions abounded: How does performance vary according to students' backgrounds? Are there particular patterns of success across tasks? What problem-solving approaches do students use and how do those affect performance?

What did NAEP learn?

Although managing equipment and training administrators requires ingenuity and painstaking effort, conducting hands-on assessment is feasible and extremely worthwhile. The school administrators, teachers, students, and consultants were all very enthusiastic. The students found the materials engaging, and the school staff and consultants were more than supportive in encouraging further use of these kinds of tasks in both instruction and assessment.

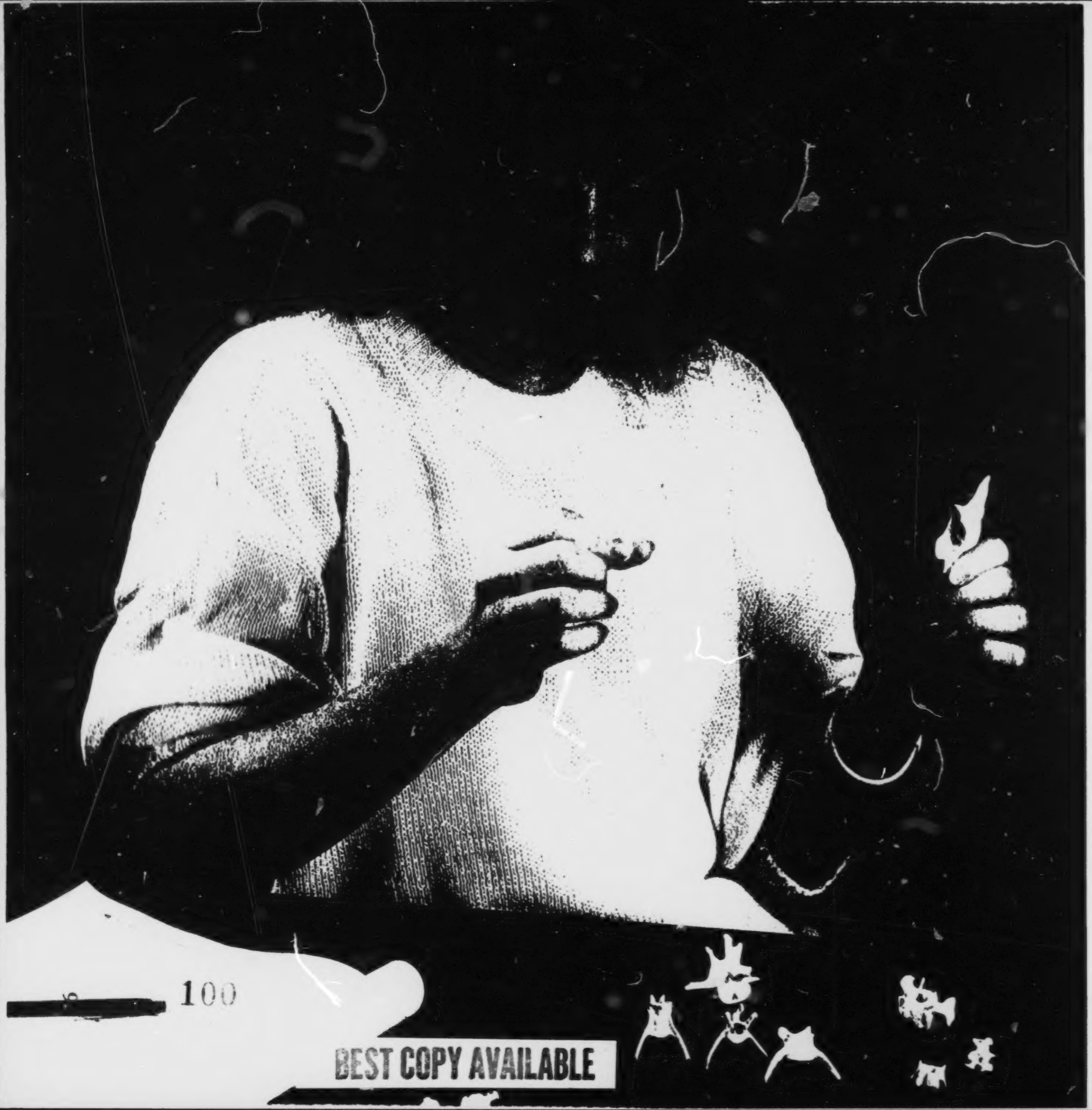
Many educators hope for systematic changes that will enable more hands-on teaching in science and mathematics classrooms. Teachers need the political, financial, and administrative support that will allow them to concentrate on developing ideas and building up the process skills necessary for students to learn to solve problems and accomplish complex tasks.

Why this manual?

In response to the interest and enthusiasm shown in the pilot study, *Learning by Doing* presents 11 tasks field-tested by NAEP. These were selected to show a range of possibilities for both classroom and assessment use. Many of the ideas underlying the hands-on tasks can be adapted to a variety of different science and mathematics concepts. In addition, such procedures as teacher demonstrations using apparatus, paper/pencil applications of some aspects of thinking tasks, and computer simulations can be integrated with hands-on experiences to ease the burden of managing students and equipment.

Each of the following illustrative tasks is identified by the thinking skills necessary for successful student performance and the administration mode used by NAEP. The presentation for each task includes a brief explanation of the activity, the student response sheet, a list of the equipment used, and one or more exemplary student responses.

Learning by Doing is adapted from *A Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics, Final Report.* This two-volume, 537-page report, which describes NAEP's project in detail and presents all 30 tasks included in the pilot study—six group activities, 20 station activities, and four complete experiments—is available from NAEP, CN 6710, Princeton, NJ 08541-6710 for $35.00 plus shipping and handling.



BEST COPY AVAILABLE

Vertebrae

Station Activity, Grades 7 and 11

Students are asked to sort a collection of small-animal vertebrae into three groups and explain how the bones in those groupings are alike. To complete this task, students need to make careful observations about the similarities and differences among the bones and to choose their categories according to sets of common characteristics.

Classifying tasks can be developed using a wide variety of objects or pictures of objects including seeds, leaves, shells, birds, fish, and flowers.

Vertebrae shows that tasks requiring classification need not be confined to younger students. Indeed, Vertebrae presented a challenge to older students, with sophisticated materials that required them to distinguish among detailed characteristics when they formed their groups.

◀ Equipment Required

Eleven bones labelled A-L as follows:

A = Lumbar dog
B = Cervical rabbit
C = Thoracic dog
D = Thoracic cat
E = Lumbar dog
F = Atlas dog
G = Cervical rabbit
H = Cervical dog
J = Lumbar rabbit
K = Thoracic rabbit
L = Lumbar rabbit

Classifying

The question with successful responses

WHAT IS THE SAME ABOUT THE BONES IN EACH GROUP?

Here's what you do:

1) Look at the collection of labelled bones. These bones are from the backbones of different animals.

Activities to Conduct

2) Put the bones into three groups. Make sure that there is something the same about all the bones in each group. You must use all the bones.

What did you find:

Record Findings

3) Write the letters of the bones in your three groups.

Group A: C, D, K

Group B: A, E, J, L

Group C: B, F, G, H

4) What is the same about the bones in each of your three groups?

Account for Findings

Group A: all have one long piece projecting; all have a hole in middle of central part

Group B: all have a central large area with hole and two long pieces projecting out

Group C: all are essentially a central structure with a hole in the middle and no long thin pieces projecting off them

(Grade 11) ▶

Wig-Wag

Station Activity, Grades 3 and 7

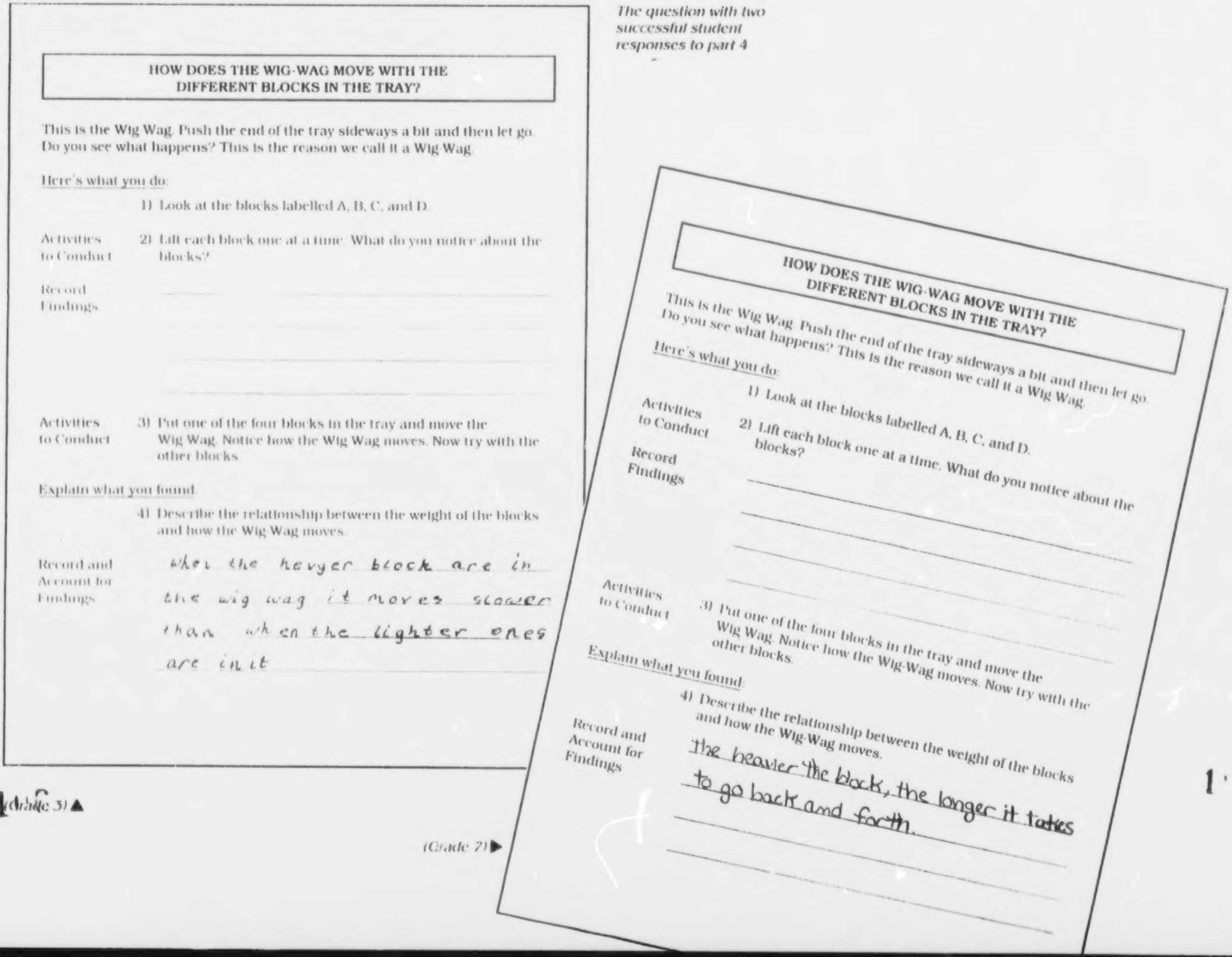
Students compare the weights of each of four blocks and observe how each individual block affects the movement of the Wig-Wag apparatus. The students then are asked to describe the relationship between the weight of each block and how the apparatus moves. To complete this task successfully, students need to carefully observe how each of the blocks affects the motion of the Wig-Wag, integrate these findings, and make generalizations about the relationship between weight and rate of movement.

◀Equipment Required

- One inertia balance
- Two large C-clamps
- One block of lead labelled A
- One block of aluminum labelled B
- One block of wood labelled C
- One block of balsa wood labelled D
- A pan scale
- A timer
- Graph paper

BEST COPY AVAILABLE

Observing and Making Inferences



The question with two successful student responses to part 4

HOW DOES THE WIG-WAG MOVE WITH THE DIFFERENT BLOCKS IN THE TRAY?

This is the Wig Wag. Push the end of the tray sideways a bit and then let go. Do you see what happens? This is the reason we call it a Wig Wag.

Here's what you do:

1) Look at the blocks labelled A, B, C, and D.

Activities to Conduct

2) Lift each block one at a time. What do you notice about the blocks?

Record Findings

Activities to Conduct

3) Put one of the four blocks in the tray and move the Wig Wag. Notice how the Wig Wag moves. Now try with the other blocks.

Explain what you found:

4) Describe the relationship between the weight of the blocks and how the Wig Wag moves.

Record and Account for Findings

whei the hevyer block are in the wig wag it moves slower than when the lighter ones are in it

(Grade 3)▲

HOW DOES THE WIG-WAG MOVE WITH THE DIFFERENT BLOCKS IN THE TRAY?

This is the Wig Wag. Push the end of the tray sideways a bit and then let go. Do you see what happens? This is the reason we call it a Wig Wag.

Here's what you do:

1) Look at the blocks labelled A, B, C, and D.

Activities to Conduct

2) Lift each block one at a time. What do you notice about the blocks?

Record Findings

Activities to Conduct

3) Put one of the four blocks in the tray and move the Wig Wag. Notice how the Wig-Wag moves. Now try with the other blocks.

Explain what you found:

4) Describe the relationship between the weight of the blocks and how the Wig-Wag moves.

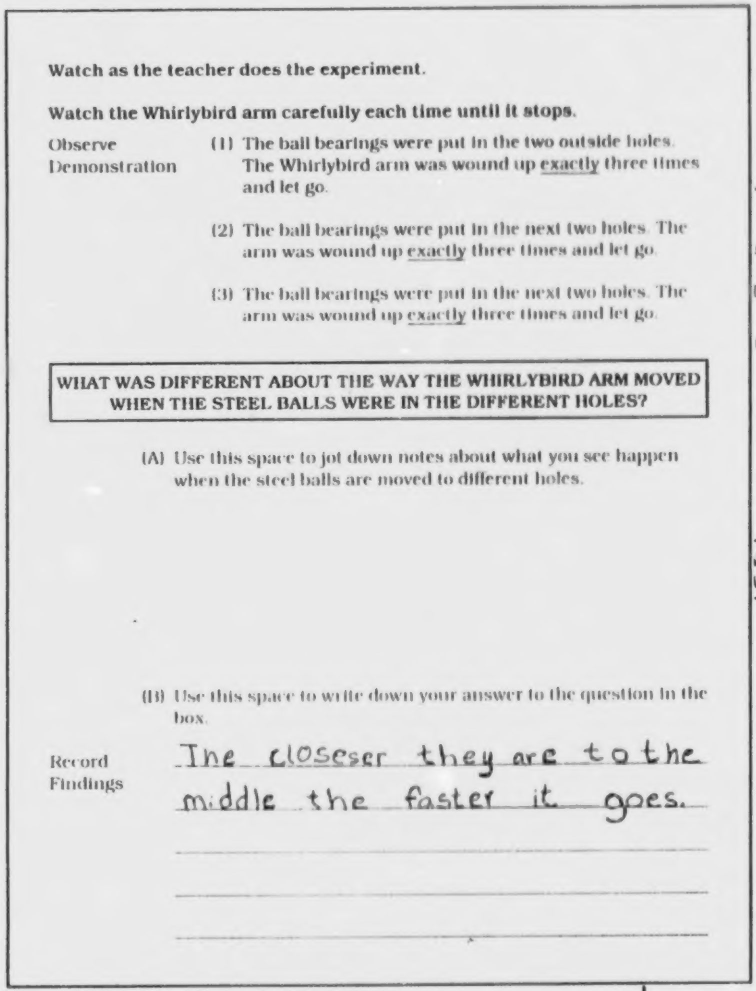
Record and Account for Findings

the heavier the block, the longer it takes to go back and forth.

(Grade 7)▶

Observing and Making Inferences

The question with two successful student responses



Watch as the teacher does the experiment.

Watch the Whirlybird arm carefully each time until it stops.

Observe Demonstration

(1) The ball bearings were put in the two outside holes. The Whirlybird arm was wound up exactly three times and let go.

(2) The ball bearings were put in the next two holes. The arm was wound up exactly three times and let go.

(3) The ball bearings were put in the next two holes. The arm was wound up exactly three times and let go.

WHAT WAS DIFFERENT ABOUT THE WAY THE WHIRLYBIRD ARM MOVED WHEN THE STEEL BALLS WERE IN THE DIFFERENT HOLES?

(A) Use this space to jot down notes about what you see happen when the steel balls are moved to different holes.

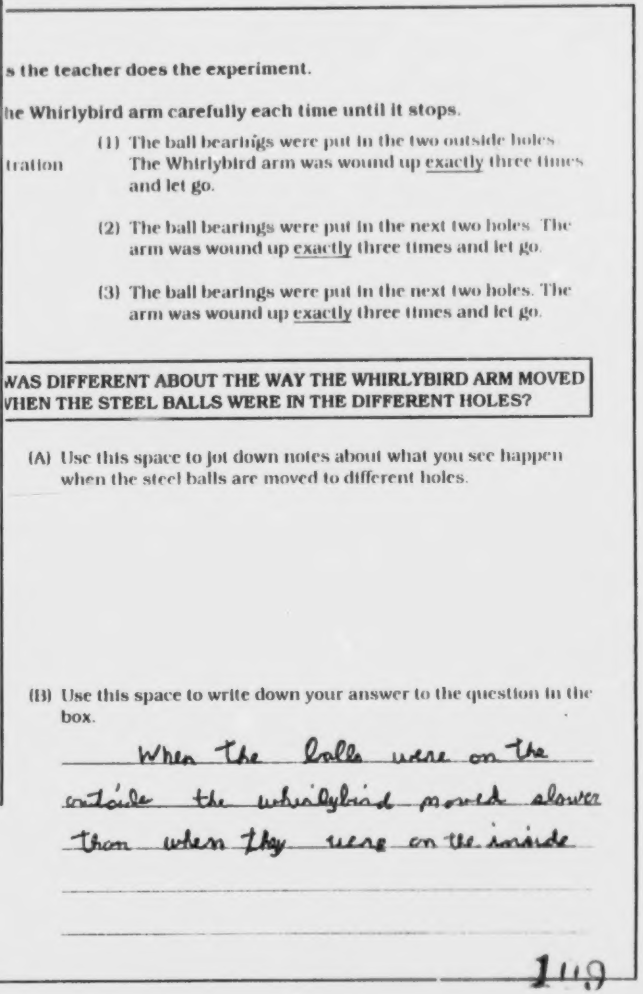
(B) Use this space to write down your answer to the question in the box.

Record Findings

The closeser they are to the middle the faster it goes.

(Grade 3) ▲

(Grade 7) ▶



s the teacher does the experiment.

he Whirlybird arm carefully each time until it stops.

tration

(1) The ball bearings were put in the two outside holes. The Whirlybird arm was wound up exactly three times and let go.

(2) The ball bearings were put in the next two holes. The arm was wound up exactly three times and let go.

(3) The ball bearings were put in the next two holes. The arm was wound up exactly three times and let go.

VAS DIFFERENT ABOUT THE WAY THE WHIRLYBIRD ARM MOVED VHEN THE STEEL BALLS WERE IN THE DIFFERENT HOLES?

(A) Use this space to jot down notes about what you see happen when the steel balls are moved to different holes.

(B) Use this space to write down your answer to the question in the box.

When the balls were on the outside the whirlybird moved slower than when they were on the inside

Whirlybird

Group Exercise, Grades 3 and 7

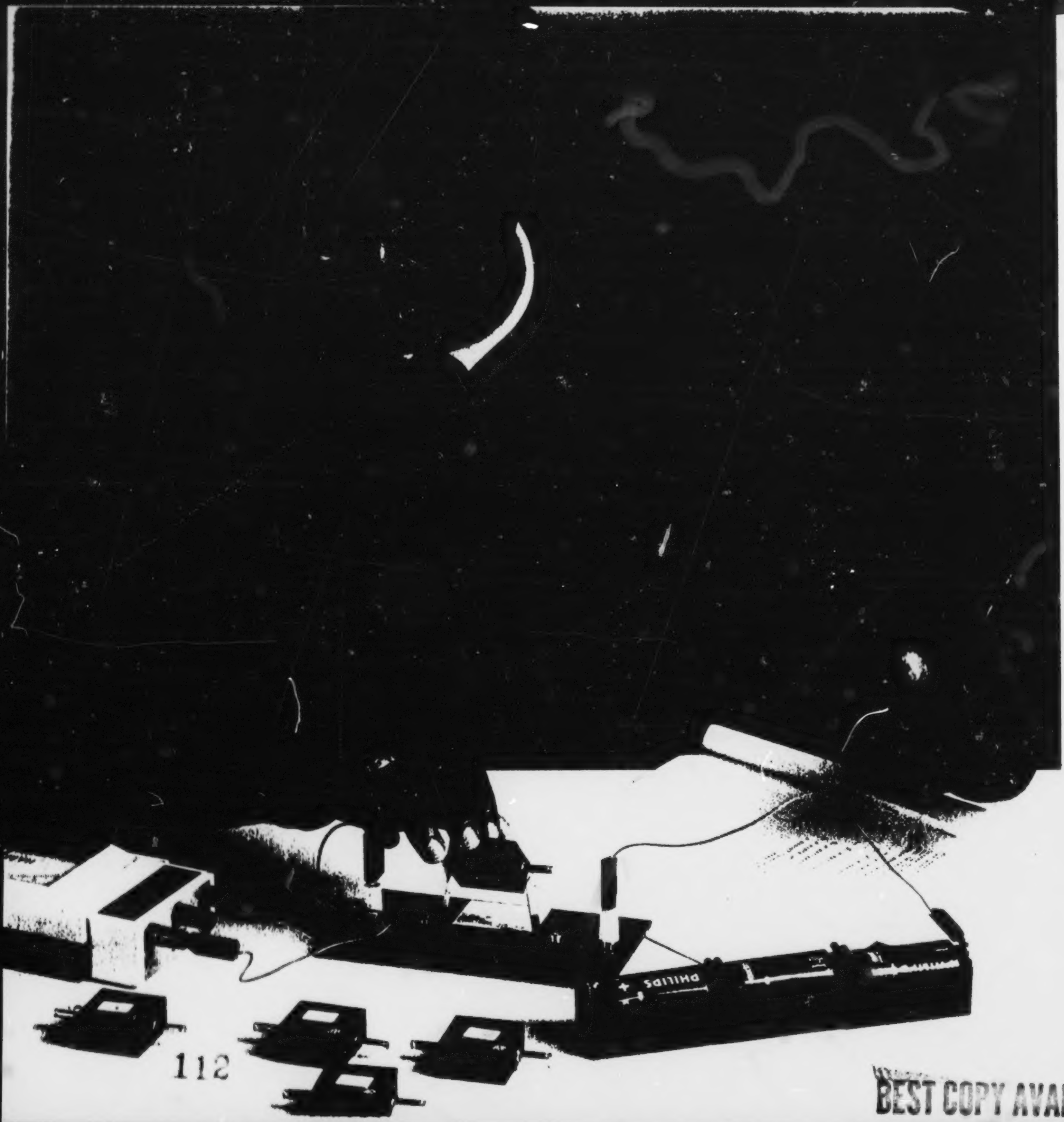
Students watch an administrator's demonstration of centrifugal force and then respond to written questions about what occurred in the demonstration. Students need to make careful observations about what happens as the administrator puts the steel balls in different holes on the Whirlybird arms, and then infer the relationship between the position of the steel balls and the speed at which the arm rotates.

Equipment Required

▶

- "Whirlybird" apparatus
- 6 ball bearings of equal mass and volume
- Spare rubber bands
- A spare ball bearing

BEST COPY AVAILABLE



Conductivity

Station Activity, Grade 11

Students identify each of five identically sealed objects by connecting the boxes that encase them, one by one, to an electric circuit. The students need to make careful observations and interpretations of what occurs as each sealed box is tested. Some knowledge of electric circuits and the conductivity of different materials is needed for this exercise.

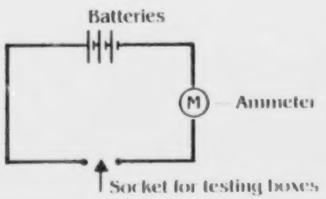
◀ Equipment Required

Five sealed black boxes labelled A-E containing the following materials:

A = a piece of copper wire
B = a resistor
C = a piece of wood
D = a diode
E = a micro relay (variable conductor)

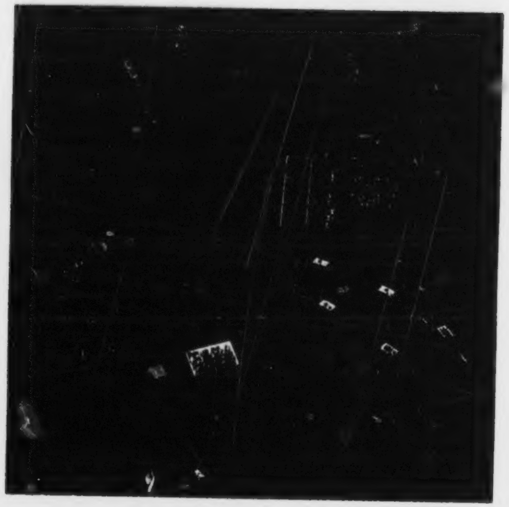
- A circuit set up with three 1.5-volt batteries in holder
- A socket for testing the boxes
- Three spare batteries

Apparatus for the circuit should be set up as shown in the diagram below.

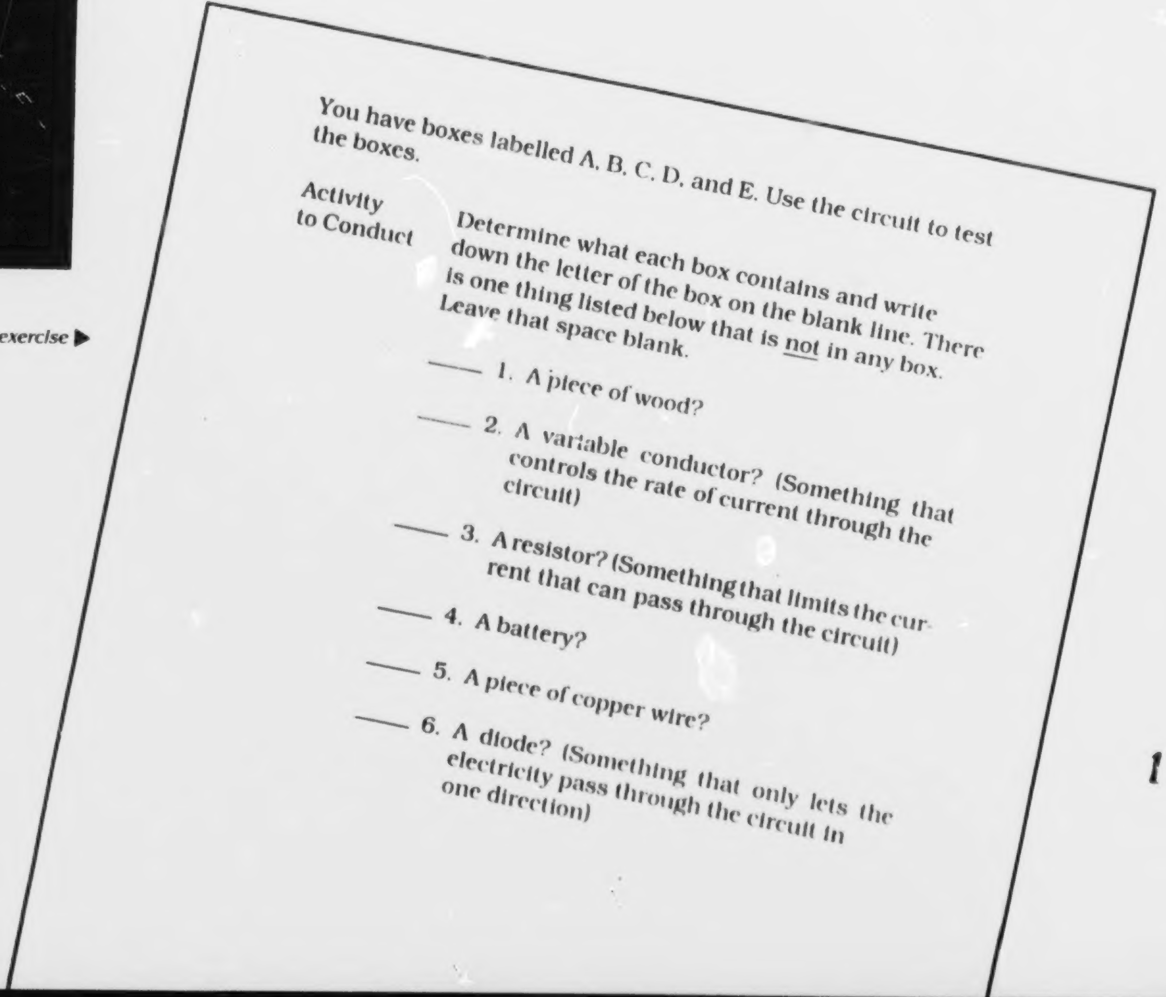


BEST COPY AVAILABLE

Observing and Making Inferences



The exercise ▶



You have boxes labelled A, B, C, D, and E. Use the circuit to test the boxes.

Activity to Conduct

Determine what each box contains and write down the letter of the box on the blank line. There is one thing listed below that is not in any box. Leave that space blank.

____ 1. A piece of wood?

____ 2. A variable conductor? (Something that controls the rate of current through the circuit)

____ 3. A resistor? (Something that limits the current that can pass through the circuit)

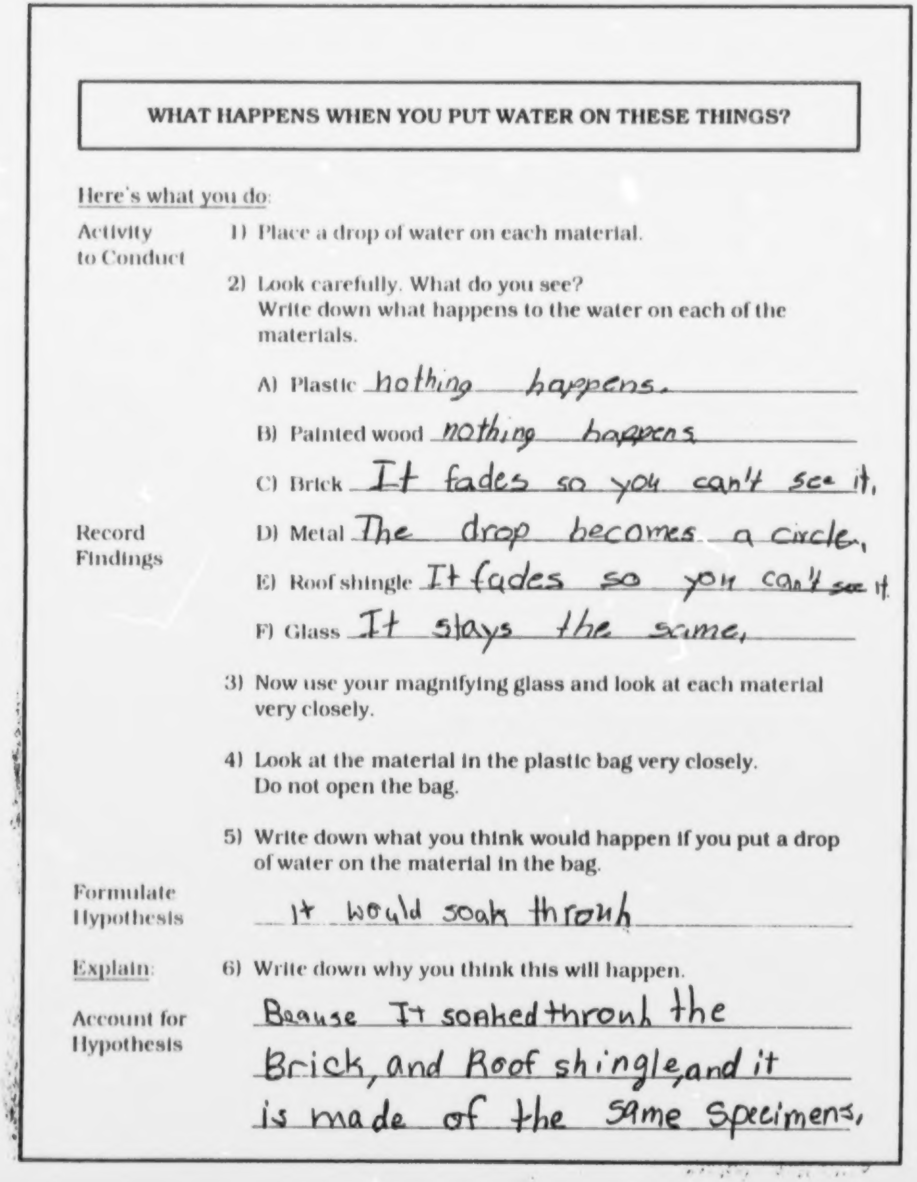
____ 4. A battery?

____ 5. A piece of copper wire?

____ 6. A diode? (Something that only lets the electricity pass through the circuit in one direction)

Formulating Hypotheses

The question with a successful student response



WHAT HAPPENS WHEN YOU PUT WATER ON THESE THINGS?

Here's what you do:

Activity to Conduct

1) Place a drop of water on each material.

2) Look carefully. What do you see?
Write down what happens to the water on each of the materials.

Record Findings

A) Plastic nothing happens.
B) Painted wood nothing happens
C) Brick It fades so you can't see it.
D) Metal The drop becomes a circle.
E) Roof shingle It fades so you can't see it.
F) Glass It stays the same.

3) Now use your magnifying glass and look at each material very closely.

4) Look at the material in the plastic bag very closely.
Do not open the bag.

5) Write down what you think would happen if you put a drop of water on the material in the bag.

Formulate Hypothesis

It would soak throuh

Explain: 6) Write down why you think this will happen.

Account for Hypothesis

Beause It soaked throuh the Brick, and Roof shingle,and it is made of the same Specimens.

◀ *(Grade 3)*

Water on Brick

Station Activity, Grades 3 and 7

Students describe what occurs when a drop of water is placed on each of seven different types of building material. The students then are asked to predict what will happen to a drop of water as it is placed on the surface of an unknown material, which is sealed in a plastic bag so that they can examine but not test it. For this exercise, students need to make careful observations, record findings, and apply what they have learned by hypothesizing what the water will do when placed on an unknown material.

Equipment Required ▶

- Eyedropper
- Small bottle filled with water
- Small, equal-sized pieces of plastic, painted wood, brick, metal, roof shingle, and glass
- An unknown material (piece of porous cinder block) in a transparent plastic bag
- A magnifying glass
- Paper
- A pencil

BEST COPY AVAILABLE

BEST COPY AVAILABLE

Double Staircase

Station Activity, Grades 7 and 11

Students are given a permanently assembled double staircase four blocks high and some loose blocks. Students first determine how many blocks are in the given staircase and then apply numerical reasoning to figure out how many blocks would be needed to build similar staircases six and 10 blocks high. Finally, students are asked to determine the mathematical relationship between a staircase of any given height and the number of blocks needed to build it.

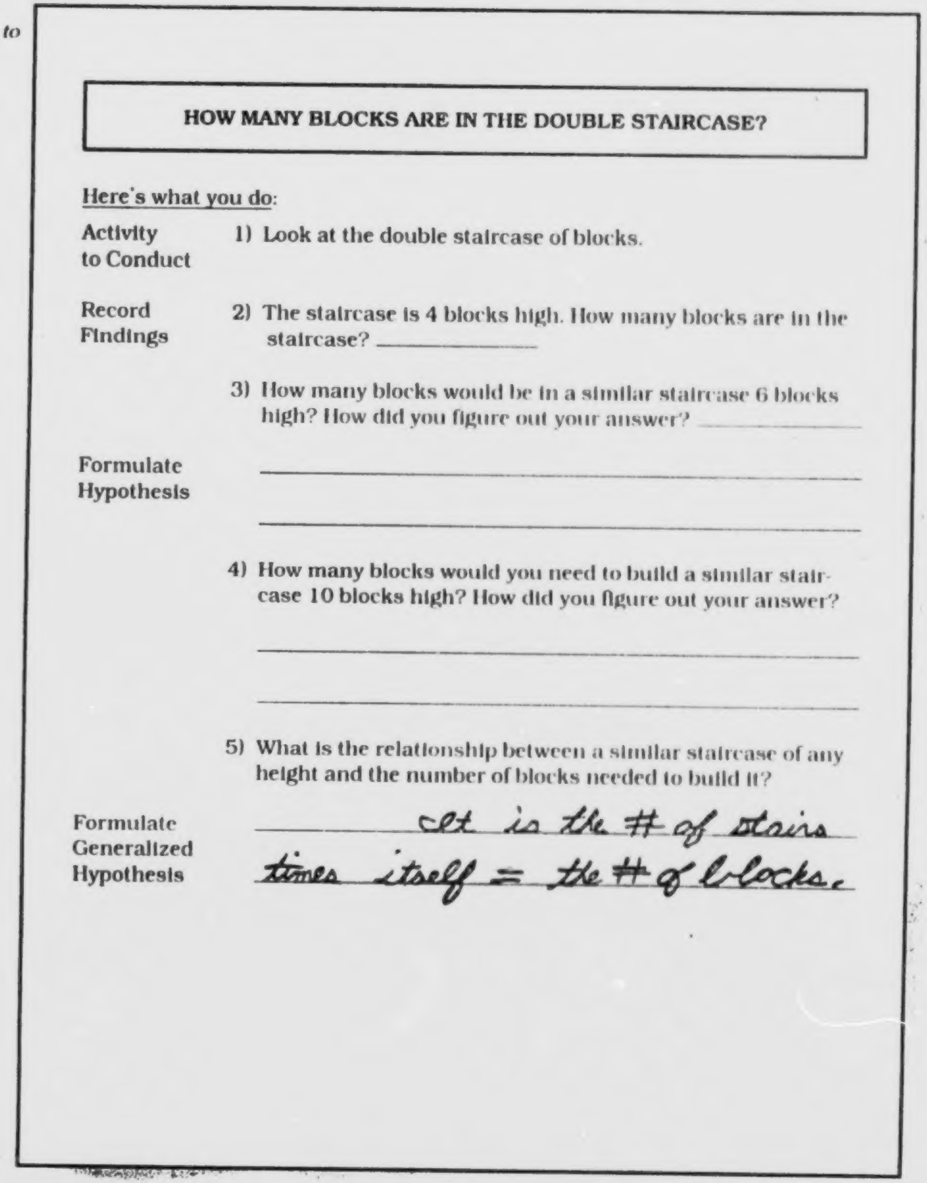
◀ Equipment Required

- Double staircase of wooden blocks that is 4 blocks high and glued to a base
- 24 loose wooden blocks that are identical to those used in the staircase
- Graph paper
- A pencil

Note: The 24 loose blocks permit a student to extend the staircase to six blocks high, but are not enough to build a staircase ten blocks high.

Formulating Hypotheses

The question with a successful response to part 5



HOW MANY BLOCKS ARE IN THE DOUBLE STAIRCASE?

Here's what you do:

Activity to Conduct 1) Look at the double staircase of blocks.

Record Findings 2) The staircase is 4 blocks high. How many blocks are in the staircase? ____________

3) How many blocks would be in a similar staircase 6 blocks high? How did you figure out your answer? ____________

Formulate Hypothesis ____________

4) How many blocks would you need to build a similar staircase 10 blocks high? How did you figure out your answer?

____

5) What is the relationship between a similar staircase of any height and the number of blocks needed to build it?

Formulate Generalized Hypothesis It is the # of stairs times itself = the # of blocks.

(Grade 7) ▶

Triathlon

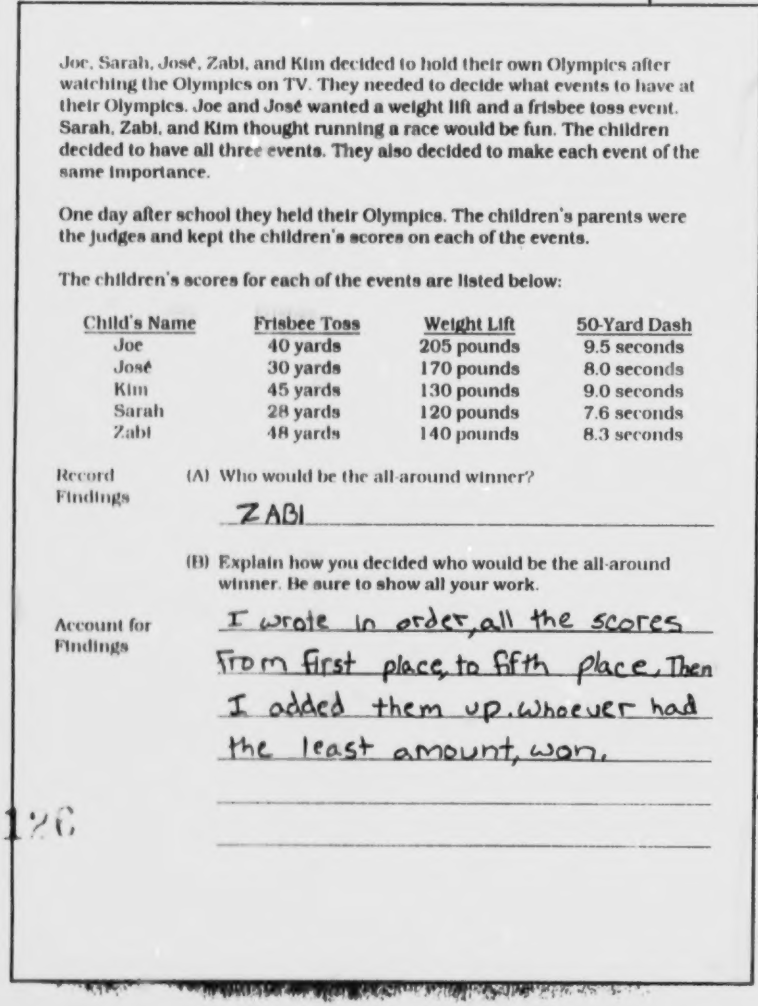
Group Activity, Grades 3, 7, and 11

Students are required by this paper and pencil task to evaluate the results of five children in three athletic events (i.e., frisbee toss, weight lift, and 50-yard dash) and decide which of the five children would be the all-around winner. Students need to devise their own approach for reviewing and interpreting the data, apply it, and explain why they selected a particular "winner." Students also need to be careful in their interpretation because lower scores in the 50-yard dash are better than higher scores, while the converse is true in the frisbee toss and weight lift.

BEST COPY AVAILABLE

Interpreting Data

The question with successful student responses



Joe, Sarah, José, Zabi, and Kim decided to hold their own Olympics after watching the Olympics on TV. They needed to decide what events to have at their Olympics. Joe and José wanted a weight lift and a frisbee toss event. Sarah, Zabi, and Kim thought running a race would be fun. The children decided to have all three events. They also decided to make each event of the same importance.

One day after school they held their Olympics. The children's parents were the judges and kept the children's scores on each of the events.

The children's scores for each of the events are listed below:

Child's Name	Frisbee Toss	Weight Lift	50-Yard Dash
Joe	40 yards	205 pounds	9.5 seconds
José	30 yards	170 pounds	8.0 seconds
Kim	45 yards	130 pounds	9.0 seconds
Sarah	28 yards	120 pounds	7.6 seconds
Zabi	48 yards	140 pounds	8.3 seconds

Record Findings

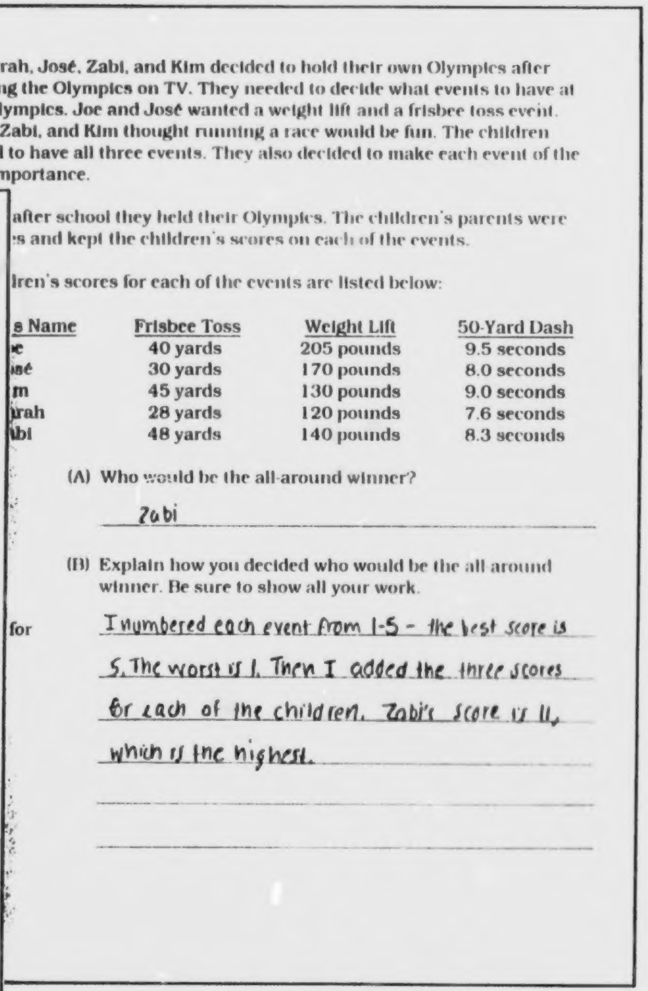
(A) Who would be the all-around winner?

ZABI

(B) Explain how you decided who would be the all-around winner. Be sure to show all your work.

Account for Findings

I wrote in order, all the scores from first place, to fifth place. Then I added them up. Whoever had the least amount, won.



Joe, Sarah, José, Zabi, and Kim decided to hold their own Olympics after watching the Olympics on TV. They needed to decide what events to have at their Olympics. Joe and José wanted a weight lift and a frisbee toss event. Sarah, Zabi, and Kim thought running a race would be fun. The children decided to have all three events. They also decided to make each event of the same importance.

after school they held their Olympics. The children's parents were
s and kept the children's scores on each of the events.

lren's scores for each of the events are listed below:

s Name	Frisbee Toss	Weight Lift	50-Yard Dash
e	40 yards	205 pounds	9.5 seconds
sé	30 yards	170 pounds	8.0 seconds
m	45 yards	130 pounds	9.0 seconds
rah	28 yards	120 pounds	7.6 seconds
bi	48 yards	140 pounds	8.3 seconds

(A) Who would be the all-around winner?

Zabi

(B) Explain how you decided who would be the all-around winner. Be sure to show all your work.

for

I numbered each event from 1-5 - the best score is 5. The worst is 1. Then I added the three scores for each of the children. Zabi's score is 11, which is the highest.

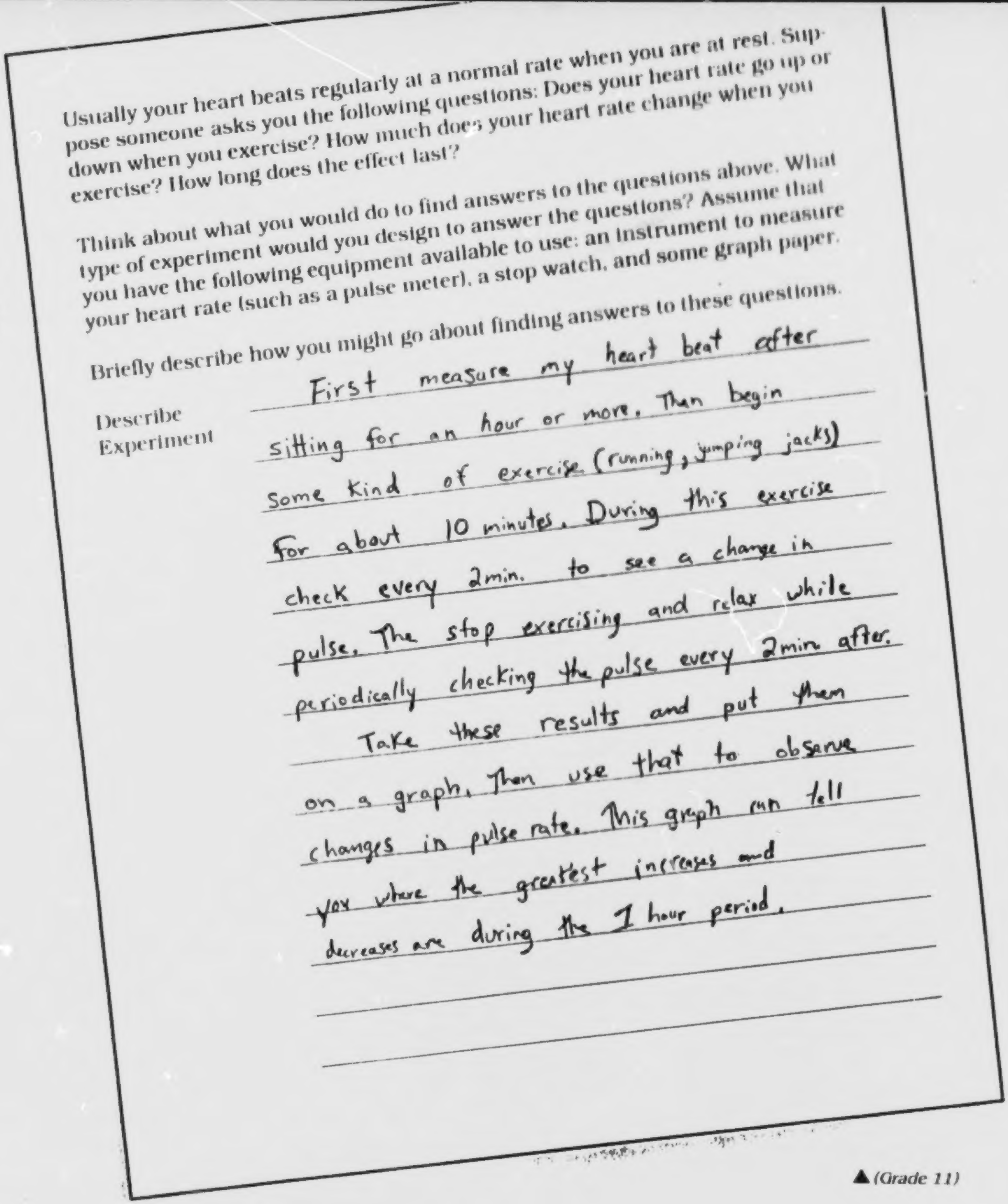
◄ *(Grade 7)*

▲ *(Grade 11)*

Heart Rate and Exercise

Grade 11

The experiment with a successful student response



Usually your heart beats regularly at a normal rate when you are at rest. Suppose someone asks you the following questions: Does your heart rate go up or down when you exercise? How much does your heart rate change when you exercise? How long does the effect last?

Think about what you would do to find answers to the questions above. What type of experiment would you design to answer the questions? Assume that you have the following equipment available to use: an instrument to measure your heart rate (such as a pulse meter), a stop watch, and some graph paper.

Briefly describe how you might go about finding answers to these questions.

Describe Experiment

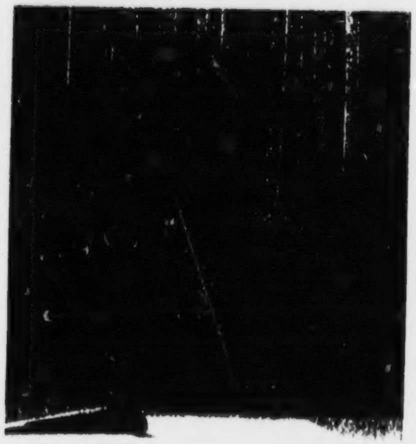
First measure my heart beat after
sitting for an hour or more. Then begin
some kind of exercise (running, jumping jacks)
for about 10 minutes. During this exercise
check every 2min. to see a change in
pulse. The stop exercising and relax while
periodically checking the pulse every 2min after.
Take these results and put them
on a graph. Then use that to observe
changes in pulse rate. This graph can tell
you where the greatest increases and
decreases are during the 1 hour period.

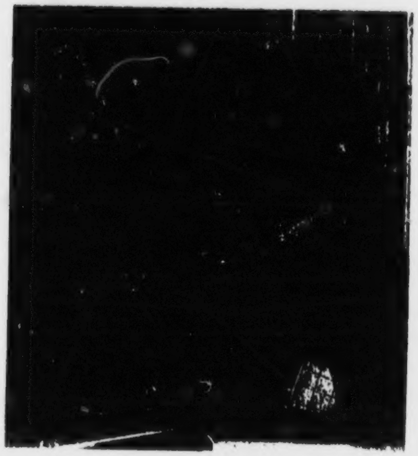
▲ *(Grade 11)*

Students design a reliable experiment to determine the effects of exercise on heart rate. In designing this experiment, students need to identify the variables to be manipulated, specify what needs to be measured, and describe how the measurements should be made to provide reliable results. This exercise is included as a prototype technique to assess students' understanding and planning of scientific investigations when actual experimentation in a classroom or assessment setting is difficult.

Designing an Experiment

Identifying Variables, Formulating Hypotheses, and Specifying Measurements





BEST COPY AVAILABLE

BEST COPY AVAILABLE

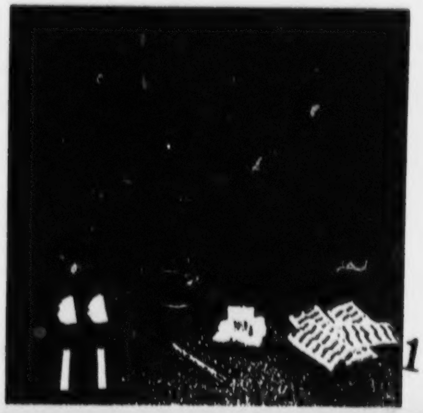
Sugar Cubes

Complete Experiment, Grade 3

Students are given laboratory equipment and asked to determine which type of sugar, granulated or cubed, dissolves faster when placed in warm water that is stirred and not stirred, respectively. To complete this investigation, students need to identify the variables to be manipulated, controlled, and measured. They also need to make reliable and accurate measurements, record their findings, and draw conclusions.

◀Equipment Required

- Six small glass beakers
- Sugar cubes in packet
- Six packages of granulated sugar, each containing the same mass of sugar as in one cube
- Hot water in thermos (50°C-60°C)
- Two stirrers
- A timer
- A graduated beaker
- A measuring cup
- A graduated cylinder
- A small ruler
- Paper towels
- Paper
- A pencil



Conducting a Complete Experiment

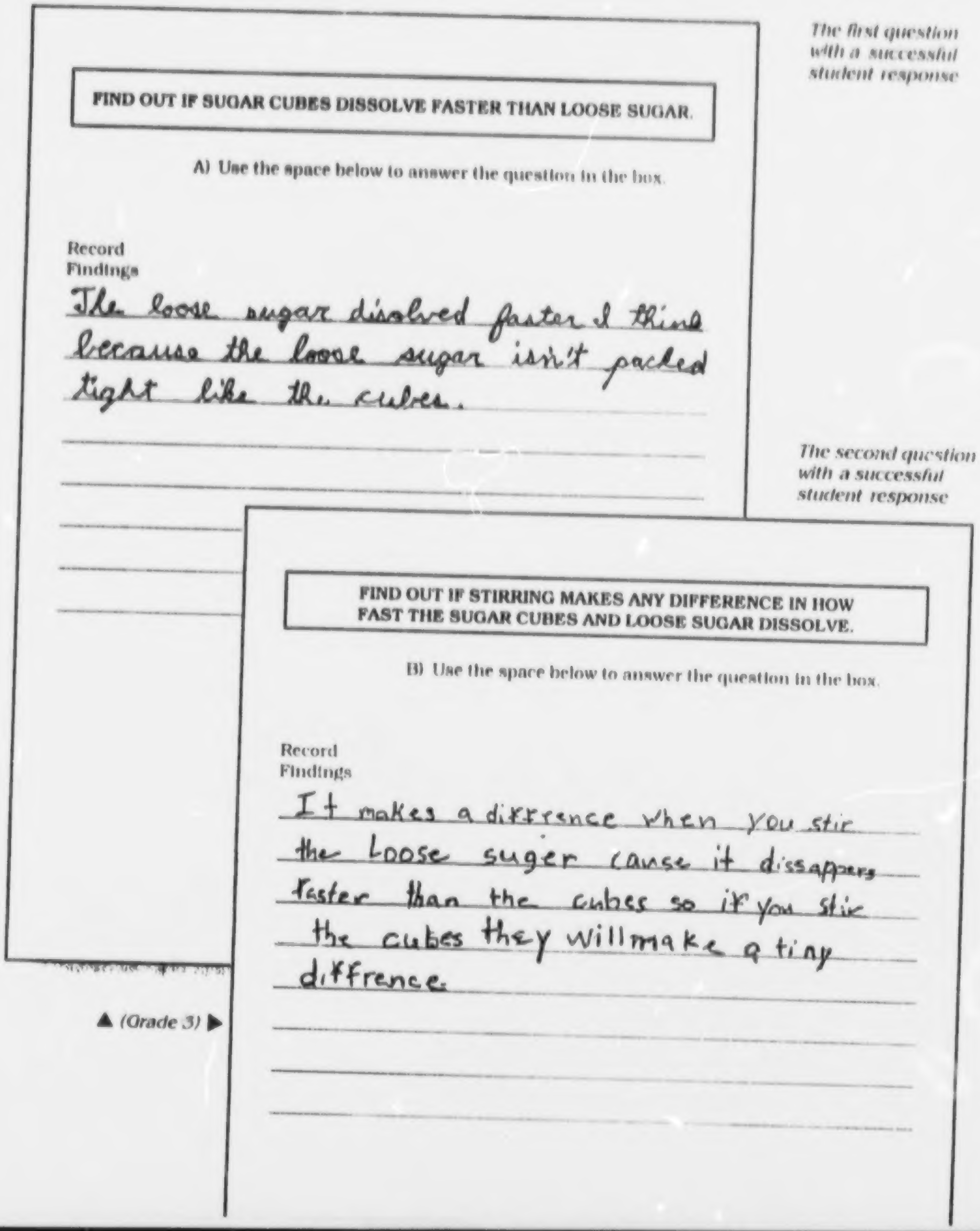
Identifying Variables, Formulating Hypotheses, Taking Measurements, Interpreting Data, and Drawing Conclusions

NAEP administrators used prepared scripts to present complete experiments to individual students. Most of the scripts contained brief background information on the problem, the problem itself, and an explanation of the equipment available to investigate it. As each student worked, her or his activities were recorded by the administrator on a detailed checklist covering students' approaches to the problem, including how they set up the experiment, manipulated the variables, and measured the outcome. The administrator encouraged students to make notes and record findings on a response sheet.

The Observation

Using detailed checklists, NAEP administrators recorded students' strategies for determining—with accurate and reliable measurements—whether loose sugar or sugar cubes dissolved at a faster rate. Successful strategies included:

- testing both types of sugar; and
- testing each by stirring and not stirring; and
- maintaining equal and/or consistent rates when stirring; and
- measuring to ensure equal amounts of sugar and equal amounts of water for each test.



The first question with a successful student response

FIND OUT IF SUGAR CUBES DISSOLVE FASTER THAN LOOSE SUGAR.

A) Use the space below to answer the question in the box.

Record Findings

The loose sugar disolved faster I think because the loose sugar isn't packed tight like the cubes.

The second question with a successful student response

FIND OUT IF STIRRING MAKES ANY DIFFERENCE IN HOW FAST THE SUGAR CUBES AND LOOSE SUGAR DISSOLVE.

B) Use the space below to answer the question in the box.

Record Findings

It makes a diffrence when you stir the Loose suger cause it dissappers faster than the cubes so if you stir the cubes they willmake a tiny diffrence.

▲ *(Grade 3)* ▶

BEST COPY AVAILABLE

Survival

Complete Experiment, Grades 7 and 11

Students are asked in this simulation task to determine which of two fabrics would keep them warmer on a mountainside on a cold, dry, windy day. As in Sugar Cubes, students need to identify the variables to be manipulated, controlled, and measured. They also need to make accurate and reliable measurements, record their findings, and draw a reasonable conclusion. However, the sophistication and quantity of the equipment call for more extensive procedures and measurements than in the other complete experiments.

◀ Equipment Required

- Five cans labelled A-E
- Two identical aluminum cans A and B
- One plastic can E with the same dimensions as A and B
- One aluminum can C that is the same height as A, B, and E but of a larger diameter
- One aluminum can D with the same diameter as A, B, and E but shorter in height
- A 110°C thermometer
- A stopwatch
- Rubber bands
- Pins
- Transparent tape
- Scissors
- An electric kettle
- Two graduated cylinders
- Sheets of blanket
- Sheets of plastic
- An electric fan
- A small ruler
- Graph paper
- A thermos
- Paper towels
- And pencils

BEST COPY AVAILABLE

Conducting a Complete Experiment

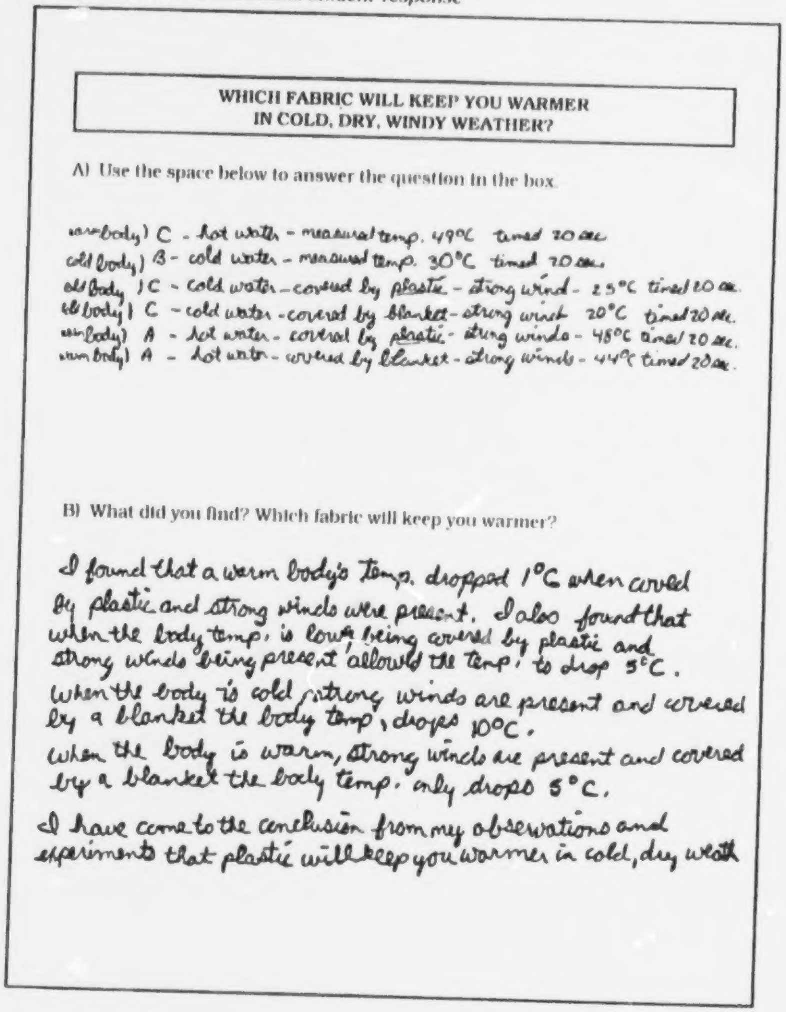
The Observation

Using detailed checklists, NAEP administrators recorded students' strategies for determining which material—blanket or plastic—would keep them warmer in cold, dry, windy weather. Students needed to be particularly careful to test both materials and use the best criteria for determining the better insulator. For example, successful strategies included:

- testing both types of materials by wrapping them around *comparable* cans of the same size, and cans that contain *equal* amounts of water at the *same* temperature;
- taking *baseline* and *final* temperature readings of the water in the cans following a *fixed period* of time OR taking a reading of the time following a *fixed temperature drop.*

As in the other complete experiments, successful investigations included accurate and reliable use of the equipment. In Survival, this would include efficient use of the stopwatch and the thermometer.

The question with a successful student response



WHICH FABRIC WILL KEEP YOU WARMER IN COLD, DRY, WINDY WEATHER?

A) Use the space below to answer the question in the box.

warm body) C - hot water - measured temp. 49°C timed 20 sec
cold body) B - cold water - measured temp. 30°C timed 20 sec.
cold body) C - cold water - covered by plastic - strong wind - 25°C timed 20 sec.
cold body) C - cold water - covered by blanket - strong wind 20°C timed 20 sec.
warm body) A - hot water - covered by plastic - strong winds - 48°C timed 20 sec.
warm body) A - hot water - covered by blanket - strong winds - 44°C timed 20 sec.

B) What did you find? Which fabric will keep you warmer?

I found that a warm body's Temp. dropped 1°C when covered by plastic and strong winds were present. I also found that when the body temp. is lower being covered by plastic and strong winds being present allowed the temp. to drop 5°C.
When the body is cold, strong winds are present and covered by a blanket the body temp. drops 10°C.
When the body is warm, strong winds are present and covered by a blanket the body temp. only drops 5°C.
I have come to the conclusion from my observations and experiments that plastic will keep you warmer in cold, dry weath

▲ (Grade 11)

NAEP's 1990 Writing Portfolio Study

Approximately 4,000 students who participated in the 1990 NAEP writing assessment, half in grade 4 and half in grade 8, were invited to participate in a special study using portfolios. We have reproduced the Table of Contents to this 188-page report, the Introduction, and "Examples of the Narrative Scoring Guide" from Chapter 2, "Evaluating the Writing."

The full report, authored by Claudia Gentile, is titled *Exploring New Methods for Collecting Students' School-based Writing*, and was issued in April 1992 by Educational Testing Service under contract with the National Center for Education Statistics.

For ordering information on this report, write:

Education Information Branch
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Ave., NW
Washington, DC 20208-5641

or call 1-800-424-1616 (in the Washington, DC metropolitan area call 202-219-1651).

NATIONAL CENTER FOR EDUCATION STATISTICS

Exploring New Methods for Collecting Students' School-based Writing

NAEP's 1990 Portfolio Study

CLAUDIA GENTILE

APRIL 1992

THE NATION'S REPORT CARD

Prepared by Educational Testing Service for the National Assessment of Educational Progress 1990 Writing Assessment under contract with the National Center for Education Statistics, Office of Educational Research and Improvement
U.S. Department of Education

For sale by the U.S. Government Printing Office
Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328
ISBN 0-16-036189-3

Table of Contents

INTRODUCTION 2
Purpose 2
Collecting Students' Writing 5
Outline of this Report 7
CHAPTER ONE Describing the Writing 8
Types of Writing 9
Audience 10
Evidence of the Use of Process Strategies 10
Evidence of the Use of Resources for Writing 11
Length of Papers and Use of Computers 12
Types of Activities 12
Summary 16
CHAPTER TWO Evaluating the Writing 18
Developing Evaluative Guides 18
The Narrative Scoring Guide 20
The Informative Scoring Guide 21
The Persuasive Scoring Guide 22
Applying the Evaluative Guides 23
Examples of the Narrative Scoring Guide 29
Examples of the Informative Scoring Guide 45
Examples of the Persuasive Scoring Guide 52
Summary of Performance Across Domains 58
Summary 60
CHAPTER THREE Comparing Methods of Assessment 62
Features of the Assessment 62
Comparing Students' Performance 64
Lessons Learned 66
Summary 71
CHAPTER FOUR Samples of Students' Writing 76
Part 1: Narrative Writing 76
Part 2: Informative Writing 109
Part 3: Persuasive Writing 141
Part 4: Poems 158
Part 5: Letters 165
Part 6: Research Reports 171
APPENDIX A Demographic Characteristics 184
APPENDIX B Students' Performance by Process Strategies 186
ACKNOWLEDGMENTS 187

I

Introduction

Purpose

In recent years, teachers nationwide have been using process approaches to writing instruction to help students become effective communicators. Many students write major texts over extended periods of time, and in many classrooms, writing instruction encompasses a range of interrelated activities that engage students in pre-writing activities, drafting, and revision.[1] As a part of this process, student writers often consult with peers, teachers, and parents.[2] The aim of these methods is to enable students to produce richer, more developed pieces of writing.

However, we face a problem when we try to assess the extent to which these efforts are successful. Traditional methods of evaluating students' writing (in particular, the timed essay test) are designed to measure a specific facet of writing ability — how well students can write on an assigned topic under timed conditions.[3] They are not designed to capture the range and depth of the writing processes in which students engage during process writing instruction programs.[4]

It is possible to emulate aspects of the process approach to writing within the context of traditional writing assessment methods. For example, the time allocated for writing can be increased, and can even be held over several days to allow for peer review and other classroom activities (e.g., New Brunswick, Canada Reading and Language Arts Multi-day Assessment Program).[5] However, holding an assessment over several days poses operational difficulties, increasing the costs and complexity of assessments.

---

[1] Janet Emig, *The Composing Processes of Twelfth Graders*. (Urbana, IL: National Council of Teachers of English, NCTE Research Report No. 13, ERIC Document No. ED 058205, 1971).

[2] Nancy Atwell, "Making the grade," in *Understanding Writing: Ways of Observing, Learning, and Teaching* (2nd edition), Thomas Newkirk and Nancy Atwell, editors, (Portsmouth, NH: Heinemann, 1988).

[3] Hunter M. Breland, Roberta Camp, Robert J. Jones, Margaret M. Morris, and Donald A. Rock, *Assessing Writing Skill*, (New York: College Entrance Examination Board, 1987).

[4] C. K. Lucas, "Toward ecological evaluation, Part 1," *The Quarterly, 10* (1), 1-3, 12-17, 1988.

[5] New Brunswick Reading and Language Arts Assessment Program, (Ministry of Education, New Brunswick, Canada, 1991).

Another way of establishing stronger connections between process writing curriculums and assessment methods is to adapt an instructional tool — writing portfolios — for assessment purposes.[6] Recently, schools, districts, and states have been exploring ways of using classroom writing portfolios to assess students' writing achievements. Using the writing students have produced as they engage in process writing programs establishes an immediate connection between the assessment and the writing process curriculum.[7] Recent efforts to adapt writing portfolios for assessment purposes can be classified into three types: the classroom portfolio, the combination portfolio, and the assessment portfolio.

The Classroom Portfolio While Classroom Portfolios differ from classroom to classroom, they usually share several key characteristics. During the school year, as part of their English/language arts classwork, students collect their written work in folders. At specific points in the term, they review their work and create a portfolio by engaging in a process of reflection, selection, and description. (e.g., New York City Portfolio Project, ARTS Propel).[8]

The reflection and selection stages are guided by a set of criteria developed by teachers and/or students, based on the writing curriculum they are following.[9] These criteria often focus on the depth of student writing (writing that demonstrates the use of process strategies and writing that shows growth over time) and on the breadth of student writing (writing that illustrates the range of activities in which students have engaged).

Often the students determine how many pieces to include in their portfolios, with a minimum of three being common practice. A central element of these portfolios is the letters or statements students write explaining their selections and how their choices meet the selection criteria. This process of reviewing and evaluating one's own writing and then articulating one's decisions is considered central to the portfolio experience because it fosters students' development as writers.[10] The classroom teachers assist students throughout this process and also evaluate the portfolios. Sometimes other students, friends, and family read and comment on students' portfolios.[11] Students may collect portfolios for part of the year, the whole year, or over their whole academic careers, for one class or all classes.

---

[6] S. Murphy and M. A. Smith. "Talking about portfolios." *The Quarterly. 12* (2). 1990.

[7] D. Galleher. "Assessment in context: Toward a national writing project model." *The Quarterly. 9.* (3). 5-7. 1987.

Robert J. Tierney, Mark A. Carter, and Laura E. Desai. *Portfolio Assessment in the Reading-Writing Classroom.* (Norwood, MA: Christopher-Gordon Publishers, Inc., 1991).

[8] Roberta Camp. "Thinking together about portfolios." *The Quarterly. 12.* (2). 8-14. 27. 1990.

Mary Fowles and Claudia Gentile. *Evaluation Report of CUNY Lehman's Writing Across the Curriculum Program.* (Princeton, NJ: Educational Testing Service. 1989).

[9] Denny P. Wolf. "Opening up assessment." *Educational Leadership. 45.* (4). 24-29. December. 1987/January. 1988.

[10] E. Winner and E. Rosenblatt. "Tracking the effects of the portfolio process: What changes and when?" *Portfolio. 1* (5). 21-26. 1989.

The Combination Portfolio The second type of portfolio assessment system uses a combination of approaches to collect writing from students (e.g., Vermont Portfolio Project).[12] In addition to asking students to assemble a portfolio from the work they have collected for their classes, students are asked to select a "best piece" and to include in their letter describing their portfolio an explanation of what makes this their best effort. Students may also be asked to complete a writing activity common to all students in a particular class or group. These three components — portfolio, best piece, and common piece — are then evaluated individually by one or more teachers and evaluative information is presented on each component, resulting in a profile of an individual student's writing achievements. Summary statements to students about their entire portfolios are also made by their classroom teacher, other teachers, and/or other students.

The Assessment Portfolio The third type of portfolio assessment system involves administering several common writing activities to students (e.g., Rhode Island Portfolio Project).[13] Committees of teachers design a series of multi-day writing activities that reflect their writing curriculum. On the same days, using the same administration procedures, the teachers have their students engage in these activities. They collect the students' work in folders and have the students review their work and write letters explaining which activity yielded the best writing and from which they learned the most. A committee of teachers then meets to score the students' responses to each activity. The result is a profile of each student's achievements relative to the common tasks. This type of portfolio differs from traditional essay assessments in that the activities are designed to match a specific school's or state's curriculum and the students' work is accomplished as part of their regular classroom activities rather than under standardized assessment conditions.

The 1990 NAEP Pilot Portfolio Study In keeping with these new developments, the National Assessment of Educational Progress (NAEP) has begun exploring alternative methods of assessing students' writing achievements — methods that focus on the writing students regularly produce as part of their classroom activities. NAEP conducted a pilot portfolio study in

---

[11] J. Flood and D. Lapp, "Reporting reading progress: A comparison portfolio for parents." *The Reading Teacher, 42,* (7), 508-514, 1989.

[12] R. P. Mills, "Portfolios capture rich array of student performance." *The School Administrator,* 8-11, 1989.

[13] Mary Fowles and Claudia Gentile, *Validity Study of the 1988 Rhode Island Third-Grade Writing Assessment.* (Princeton, NJ: Educational Testing Service, 1989).

1990 in order to explore the feasibility of conducting large-scale assessments using school-based writing. The main purposes of this pilot study were: (1) to explore procedures for collecting classroom-based writing from students around the country; (2) to develop methods for describing and classifying the variety of writing submitted; and (3) to create general scoring guides that could be applied across papers written in response to a variety of prompts or activities.

To this end, a nationally representative subgroup of the fourth and eighth graders who participated in NAEP's 1990 writing trend assessment was asked to work with their teachers and submit one piece of writing that they considered to be a sample of their best writing efforts. The goal was to create a "Nation's Portfolio" — a compilation of the best writing produced by fourth and eighth graders in classrooms across the country.

NAEP analyzed and summarized these samples of writing along with teachers' descriptions of the assignments that produced them. In addition, NAEP compared students' school-based writings to their responses on the 1990 NAEP writing assessment to examine relationships between these two modes of assessment. This report describes the procedures used to collect, describe, and evaluate the school-based writing in this special pilot study.

The 1990 writing assessment was a trend assessment — prompts that had been developed for the 1984 assessment, and readministered in 1988, were also given in 1990 in order to measure changes in students' writing achievements across the six-year period. In 1992, NAEP will continue the writing trend assessment, as well as conduct a new writing assessment comprised of informative, narrative, and persuasive writing prompts developed specifically for the 1992 assessment. While the trend writing assessment has not changed since 1984, the new 1992 writing assessment reflects recent developments in the field of writing instruction and assessment. For example, the time allocated for writing has been expanded to 25- and 50-minute periods. Also, a planning page has been included after each prompt, to encourage students to reflect and plan their responses to the topics. The 1992 assessment will also include a revised and expanded version of the 1990 pilot portfolio study and participants will be selected from among those students taking the new regular writing assessment.

Collecting Students' Writing

The Participants Approximately 4,000 students who participated in the 1990 NAEP writing assessment — 2,000 students at grade 4 and another 2,000 students at grade 8 — were invited to participate in the special portfolio study. Based on traditional NAEP sampling procedures, this group would have been a nationally representative sample of the nation's fourth and eighth graders.

However, only 55 percent (1,110 students) of the fourth graders and 54 percent (1,101 students) of the eighth graders and/or their teachers accepted this invitation. While these response rates provided enough papers to permit an analysis of the writing submitted on a pilot basis, as statistical samples they were too small to make generalizations about all of the nation's fourth and eighth graders' writing performances.

While the participants did not represent a national sample of students, they were from all of the major geographic regions and from various types of communities, including rural, suburban, and inner city. They represented a variety of racial/ethnic backgrounds as well as a balance between males and females (see Appendix A for details on the demographic characteristics of the participants).

Compared with the entire group of students who participated in the 1990 NAEP writing assessment, the participants of this study differed in some respects. Slightly higher percentages of the portfolio pilot study participants:

- were above the modal ages of the sample (ages 9 and 13),
- attended schools in advantaged urban communities, reported having higher grades,
- reported having a greater number of reading materials at home, and
- received slightly higher scores on the NAEP writing assessment tasks.

When considering the data from this pilot study, it is important to keep in mind that the students who participated appear to be somewhat older, higher achieving, and more advantaged than the larger population of students assessed by NAEP in 1990.

The Procedures In the spring of 1990, at the time of the NAEP writing assessment, the English/language arts teachers of participating students were asked to help several of their students choose a sample of their own *best* writing from the work the students had completed so far in the 1989-90 school year. No more than 10 students from any given class were selected to participate. Teachers were asked to encourage their students to choose pieces that had involved the use of writing process strategies (such as revising successive drafts, using reference sources, consulting with others about writing). NAEP also asked teachers to attach a description of the activities that generated the students' writing and to comment on any process strategies the students used to produce their writing.

Teachers then submitted their students' writing to NAEP, along with a copy or description of the activities that generated the writing and any available drafts or prewriting samples. These pieces were used to create two national portfolios or collections of students' classroom writing — one containing the writings of fourth graders and the other containing the writings of eighth graders.

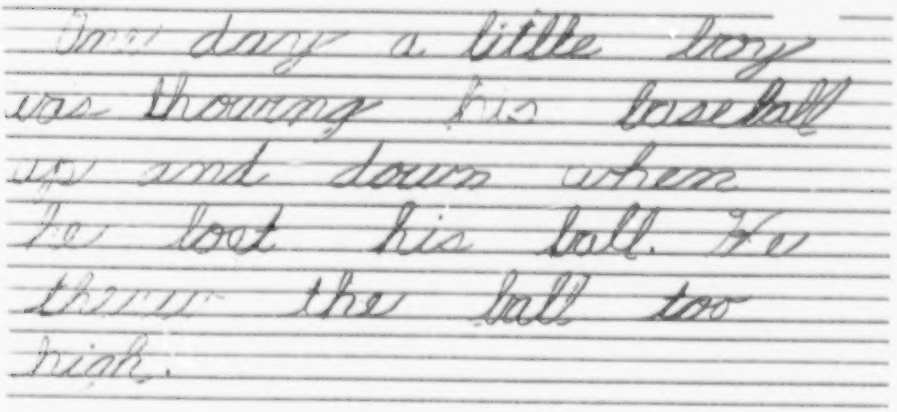
Unfortunately, due to the complex procedures NAEP employs to select students to participate in its assessments, we were unable to inform teachers at an early date which of their students would be participating in this study, with some teachers receiving only several days' notice. Thus, for the pilot, teachers and students did not have much time to review the students' writing and select best pieces. Based on this experience, a procedure for giving teachers more advance notice of the upcoming portfolio assessment was developed for the 1992 NAEP Portfolio Study. It is hoped that, by giving the participating teachers in 1992 several months' notice, the 1992 results will be representative.

Outline of this Report

This report is divided into four sections. Chapter One describes the writing received from the students and information from participating teachers about the activities that generated the writing. Chapter Two explains the procedures used to evaluate the writing students submitted as well as the results of this evaluation. Chapter Three compares the results of the NAEP 1990 writing assessment with the analysis of participants' school-based writing samples and summarizes the lessons learned from this portfolio study. The last chapter contains a set of sample papers, further illustrating how the evaluative guides can be applied and presenting a sense of the range and depth of writing we received from participating students.

Examples of the Narrative Scoring Guide

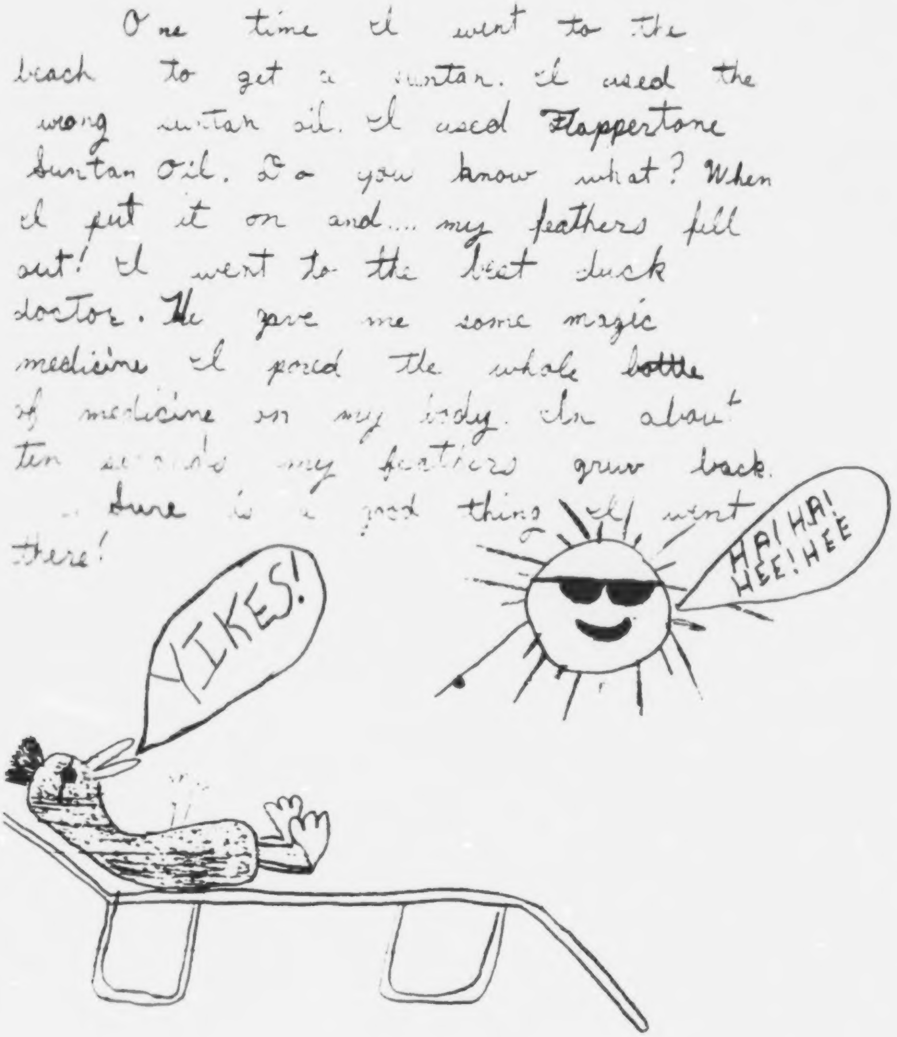
Event Description (score of 1) Papers classified as event descriptions tell about one event. Basically, they say, "such and such happened." Some of the papers in this category give details about the setting and so appear to be more elaborate stories. However, they end with a description of a single event, rather than a series of events. The paper below, written by a fourth grader, is an example of a simple *Event Description*.



One day a little boy
was throwing his baseball
up and down when
he lost his ball. He
threw the ball too
high.

Undeveloped Story (score of 2) Papers classified as *Undeveloped Stories* tell about a series of events. Basically, they say, "one day this happened, then something else happened, and then another thing happened." However, the events, as well as the setting and characters, are only briefly described. The writers give very few details about each event: the story is a listing of related events.

These stories are similar to front-page newspaper reports, where the basic facts of a story are reported (who, what, when, where) but few details about why events happened are presented. For example, in the paper below, the fourth-grade writer uses one sentence to describe each event.

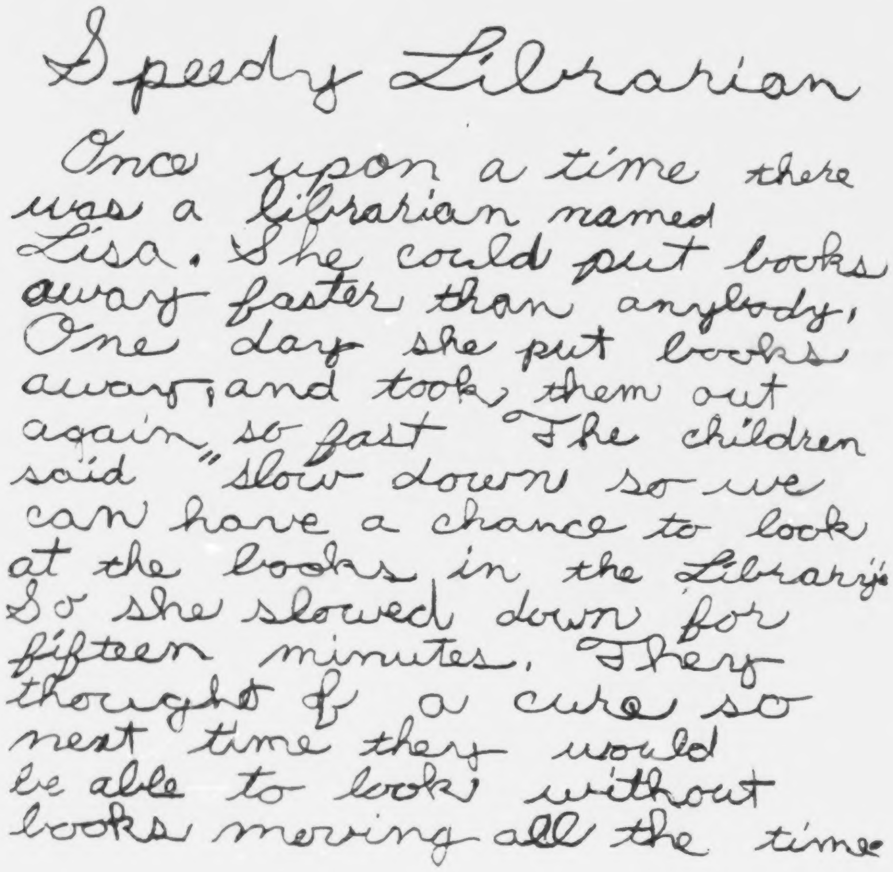


One time I went to the
beach to get a suntan. I used the
wrong suntan oil. I used Flappertone
Suntan Oil. Do you know what? When
I put it on and.... my feathers fell
out! I went to the best duck
doctor. He gave me some magic
medicine I pored the whole bottle
of medicine on my body. In about
ten seconds my feathers grew back.
Sure is a good thing I went
there!

YIKES!
HA! HA! HEE! HEE

BEST COPY AVAILABLE

Basic Story (score of 3) In papers classified as *Basic Stories*, the writers go one step beyond a simple listing of related events. One aspect of the story (the events, the characters' goals, or the setting) is somewhat developed. However, these stories lack a sense of cohesion and completeness. Events may be presented out of sequence, some aspect of the story may be confusing due to problems with syntax, or a key event may be unclear. For example, in the paper below, the fourth-grade writer describes a series of events and, at the beginning, develops a problem in some detail (a librarian who puts books away too quickly). However, the resolution to the problem, although humorous, is not well developed.



Speedy Librarian

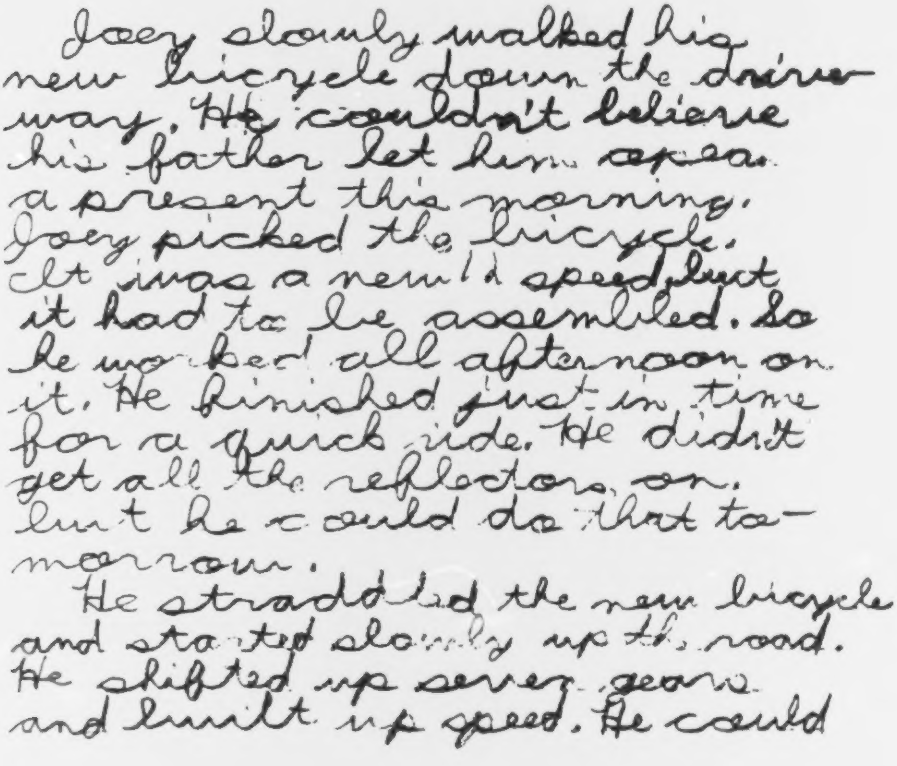
Once upon a time there was a librarian named Lisa. She could put books away faster than anybody, One day she put books away, and took them out again, so fast The children said "slow down so we can have a chance to look at the books in the Library. So she slowed down for fifteen minutes. They thought of a cure so next time they would be able to look without books moving all the time

"The next time we go to library, we will chain her to the seat for a half hour or until we're done," the children laughed! They chained her to the seat right when they arrived. Every time they went they chained her down. She was finally cured.

THE
end

Extended Story (score of 4) *Extended Stories* go beyond *Basic Stories* in that many of the events in these stories are elaborated to some degree. This degree of development gives a sense of a sequence of distinct story episodes. Details are given about the setting, the characters' goals, problems to be solved, and the key events. Yet, these stories may be somewhat incomplete in that the characters' goals may be left unresolved or the problem posed in the story's opening never solved. The ending may not match the beginning or the story's ending may be inconsistent with the internal logic established throughout the rest of the story. Or, as in the example below (written by an eighth grader), they may be very satisfying, yet not elaborately developed.

It is important to note that, while *Extended Stories* are not as elaborated or complex as are *Developed Stories* and *Elaborated Stories*, they are successful stories — all of the key story elements and events are clearly presented. They are the simplest type of complete story on this scale.



Joey slowly walked his new bicycle down the driveway. He couldn't believe his father let him open a present this morning. Joey picked the bicycle. It was a new 12 speed, but it had to be assembled. So he worked all afternoon on it. He finished just in time for a quick ride. He didn't get all the reflectors on, but he could do that tomorrow.

He straddled the new bicycle and started slowly up the road. He shifted up several gears and built up speed. He could hear the loud clicking of the rear tire. He got to a steep hill and rode down the incline. He heard the clicking grow to a steady buzz as he coasted faster and faster.

He didn't realize he was so close to the highway. He saw a pair of headlights headed toward him from his left. He locked his brakes. The driver in the car did the same, but Joey still slid out onto the highway. He felt a tremendous pain to his leg. Then another to his head. His vision faded away. He could hear people shouting and talking. Then his hearing started to fade. The last thing he could hear before he lost consciousness was the slow soft clicking of a bicycle tire

Developed Story (score of 5) *Developed Stories* describe a sequence of episodes in which almost all of the events and story elements are somewhat elaborated. Yet, one aspect of these stories is not well developed, such as the ending or a crucial event. In the example below (written by an eighth grader), each episode is somewhat developed, but could be further elaborated.

The lights were very dim in the reform school when the girls lay in there beds.

Lisa a girl at the school was very afraid of dark, tight places. She would scream and cry if someone locked her in a small, dark room.

She had told stories of her step-mother locking her in an icebox for a punishment.

"I tried not to breath," Lisa said. "I didn't want the air to get too low but after a while I had to breath".

Lisa looked like a nice girl. She had long blonde hair, stood about 5'2" and weighed about 115 pounds.

She had been an orphan for ten years until a lady came into the orphanage and took her home. The lady soon adepted her.

After a while the lady got tired of her so she put her in a reform school.

It was night time at the reform school and the girls were tired. Mrs. Bertha came in and shut the lights off and made the girls lay down.

A half hour later the reform school grew quiet. The girls had fallen asleep.

Lisa began to toss and turn. A vision of her step mother came into her mind.

"Lisa ~~get up~~ you have been very bad lately you have to be punished, her stepmother said quietly.

No, No! Lisa shouted still half asleep.

Lisa ran out through the hallway until she bumped into a huge, black door that had a rusty, half broken sign on it. Lisa apened the door as fast as she could.

When she got the door open she ran in the room.

The room was dark and cold. It had the smell of damp walls allmost like that af a very old dungeon.

When she got the door shut she locked it. After she locked the door she turned around and faced it.

The vision of her stepmother walked through the door, and went towards her slowly.

Lisa walking backwards tripped and fell into a coffin! The coffin door shut and locked.

"Let me out," Lisa screamed!"

The air inside af the coffin was getting very low. She was getting weaker and weaked every second.

At the distance she could hear a faint evil laugh. A laugh that she did not recognise, but had heard before.

Elaborated Story (score of 6) No papers were considered to be *Elaborated Stories*. To be classified as elaborated, stories had to present a sequence of episodes in which almost all of the events and story elements were well developed. Goals or problems introduced in the beginning were well resolved by the end, characters' motives were well developed, and the entire story was a cohesive, unified whole.

In the example below, the eighth-grade writer of "The Black Rose" retells the plot of a Halloween movie. In it, the writer effectively presents each episode, leading to a spine-tingling ending. The only discordant note is the occasional switching of narrative voice between first person and third person. A revising of this story that included a consistent use of narrative voice would make this an example of an *Elaborated Story*. (As is, this story received a score of 5.)

The black Rose

The night was uery stormy, the wind shook the trees like a limp doll. The rain pounded on the windows like a drum. Tonight, the night of September, Friday the 13th, a thin young girl of about 17 was getting ready to babysit. Her name was Bethany Miller. She had been a steady babysiter since about last year, there for she is saving money to by a new car. So far she had saved $2,100.

Bethany's mother and father were out at a movie. Suddenly the phone rang, Beth sprinted down the stairs to catch the phone. Beth picked up the phone to the voice of Mrs. Perkins, the woman, She was to babysit for.

"Bethany, I'm a afraid that I don't need a babysitter tonight. John and I decided not to go out because of the storm."

"Well alright, thats fine." and then our conversation ended. Actually I was kind of glad that I didn't have to baby sit, for the reason that I didn't like to go out on stormy nights, but rather curl up and read a book or watch t.v.

Beth decided to see just how bad the storm was. So she walked over to the window. The rain was so wicked that it seemed sheild Beth from the outside.

Just after Beth turned on the t.v. and sat down, the phone rang agian. Beth took her time answering the phone. It was Beth's mom.

"Honey, I'm afraid your father and I won't be home tonight. The bridge we take to get back home got flooded. Now I have a feeling the power may shut off. Trees are falling every where, so get some matches and candles and the flashlight down celler, and you might as well get the radio." And after my mom gave me a few more helpful hints, among many other things our conversation finally ended.

Chapt. 2

After I had gathered the matches, candles and radio, I went down to the cellar to get the flashlight. The cellar smelt damp. In a far corner you could hear the water from the storm trickling. As I started my way back up stairs, the lights shut off. I turned on my flashlight, then I started to hear tapping on the windows and knocking on the door, my heart started to beat fast, then as I walked up the stairs I convinced my self that it was probably just a tree nocking on everything.

Now that the power had gone out, I figured I'd turn on the radio for some comfort. Just as I was turning it on I caught a news clip about a murderer who had escaped from jail. His name was John Henry. The name caught my attention. That man was the man I heard about 2 year ago that killed 40 nurses in Baton Rouge. He was known to leave a black rose by each dead victim after he had slit there throat.

Than again the phone rang, I was kind of glad. I would like to talk to someone. As I picked up the phone and repeatedly called hello. I could hear nothing but deep breathing. Than the person on the other end hung up.

After I hung up, I was very upset and scared. Than the phone rang again, and I decided that it was the same person whom the first time thought he had the wrong number.

"Hello", I said

"Are you alone". a distant scratchy voice came over the phone.

"Who is this". I screamed

"I'm watching you". the voice said. Then I heard the voice shut up as the phone got hung up.

I trembled with fright. "Who was that"! I asked myself "Mabey it was just a kid playing", but I didnt care, I was still scared stiff. But hey, any way how could any one see me when its dark and you can't see through the windows anyway.

Even though Beth had thought of many things that it could be, she still was scared.

As Beth was curled up on the couch, with a book in one hand and and the flashlight in the other, she was jumped by a noise. As her ears perked up to listen, she could than make out the noise; it was a scratching noise coming from the cellar, almost like a cat pawing at the door, but she didn't have a cat, then it stopped. Beth was terrified with fear. She wanted to call her mother but she didn't dare move from her secured position. While Beth sunk into the blue cushions of her couch. She now heard footsteps upstairs heavy, loud footsteps that seemed to make the ceiling shake. Beth wasn't sure if this was her imagination or not. After the footsteps had stopped. The phone rang, forgetting the other phone calls Beth jumped up to answer the phone. This just had to be her mom calling. She thought with extreme excitement.

"Mom", Beth questioned

"I'm still watching you!" a muffled deep voice replied.

With out any words Beth hung up the phone and started to scream and cry. Beth could no longer ignore the phone calls and noises, she finally made up her mind to call the police, but when she picked up the phone she heard deep breathing once more.

"Hello, Hello", she stuttered

"I'm in the house!"

With those words Beth could no longer take it; she once more tried the phone. this time she got through, but just as she was finishing her name, the line got cut off. Then she turned around and a tall man with a black mask and black outfit was standing holding the phone line, then he dropped the line and held a terribly big butcher knife over his head and...

Chapt 3.

The storm had cleared and Bethis parents were ariving home. They opened the door and dropped there coats. When Beths mom looked over toward the Kitchen, she saw a trail of blood, She screamed in horror, Beth's father ran over to where he saw his daughter lying in a puddle of blood with a limp black rose on her chest,

Performance Assessment
An International Experiment

While the extensive use of paper and pencil tests in the main International Assessment of Educational Progress (IAEP) assessments made it possible to achieve good coverage of the knowledge and skills that could be assessed with such instruments, experience in the United Kingdom had demonstrated that some types of performance assessment were feasible in national surveys of student attainment. Given this experience and the desirability of extending the curriculum coverage in IAEP, a limited, optional component of performance assessment was included in the 1991 survey. The assessment was developed for 13-year-olds only and included mathematics and science tasks to enable IAEP participants to experiment with performance assessment in an international context. The U.S. did not participate.

The pages reproduced here show the science tasks used, along with the performance of the participating countries.

Performance Assessment: An International Experiment was written by Brian McLean Semple, The Scottish Office, Education Department, and published by the International Assessment of Educational Progress, Educational Testing Service, Report No. 22-CAEP-06, July 1992.

Copies of the report can be ordered from:

Center for the Assessment of Educational Progress
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-0001

The Center for the Assessment of Educational Progress (CAEP) is a division of Educational Testing Service devoted to innovative approaches to the measurement and evaluation of educational progress. The present core activity of CAEP is the administration of the National Assessment of Educational Progress (NAEP), under contract from the U.S. Department of Education. CAEP also carries out related activities, including the International Assessment of Educational Progress (IAEP), state assessments, and special studies.

PERFORMANCE ASSESSMENT:

AN INTERNATIONAL EXPERIMENT

Brian McLean Semple
The Scottish Office
Education Department

Prepared for the National Center for Education Statistics
U.S. Department of Education and the National Science Foundation

July 1992 **Report No. 22-CAEP-06**

The International Assessment of Educational Progress ETS IAEP

EDUCATIONAL TESTING SERVICE

Educational Testing Service (ETS) is a private, nonprofit corporation devoted to measurement and research, primarily in the field of education. It was founded in 1947 by the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board.

The Center for the Assessment of Educational Progress (CAEP) is a division of ETS devoted to innovative approaches to the measurement and evaluation of educational progress. The present core activity of CAEP is the administration of the National Assessment of Educational Progress (NAEP), under contract from the U.S. Department of Education. CAEP also carries out related activities, including the International Assessment of Educational Progress (IAEP), state assessments, and special studies such as the National Science Foundation-supported Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics.

The work upon which this publication is based was performed pursuant to Grant No. SDE-8955070 of the National Science Foundation, and additional funding was provided by the U.S. Department of Education through Interagency Agreement No. IAD-91-0222. Supplementary funds were provided by the Carnegie Corporation of New York.

This report, No. 22-CAEP-06, can be ordered from the Center for the Assessment of Educational Progress at Educational Testing Service, Rosedale Road, Princeton, New Jersey 08541-0001.

Library of Congress Card Number: 92-71732

ISBN: 0-88685-127-0

Educational Testing Service is an equal opportunity/affirmative action employer.

Educational Testing Service, ETS, and ETS® are registered trademarks of Educational Testing Service.

BEST COPY AVAILABLE

Science Tasks

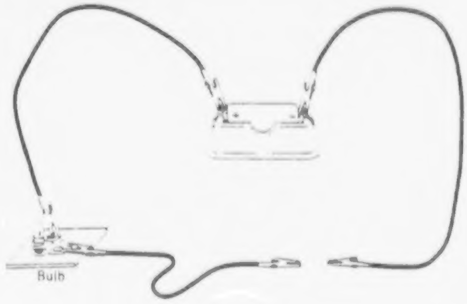
LIGHT-UP

Task Descriptor

To categorize objects according to their electrical conductivity by completing an electrical circuit; to explain why some objects enable a bulb to light; to predict whether an object in a sealed container would enable the bulb to light and to explain why (or why not).

Equipment/Material

An electrical circuit with a bulb and a gap with two contacts which could be bridged. Five objects as follows: wood strip, plastic strip, nail, foil strip and cardboard strip. Also an object (piece of copper wire) in a sealed, clear plastic box.



Student Instructions

Complete the circuit using the five objects in turn. List those objects that enable the bulb to light and explain why. Say whether you think object X would enable the bulb to light and explain why.

Scoring Scheme

Credit was given for identifying the nail and foil strip as conductors and for giving an explanation mentioning one of the following or its equivalent: objects conduct electricity, allow electricity/charge to pass, complete the circuit, are metal. Also credit was given for saying object X would enable the bulb to light and for giving an explanation as above.

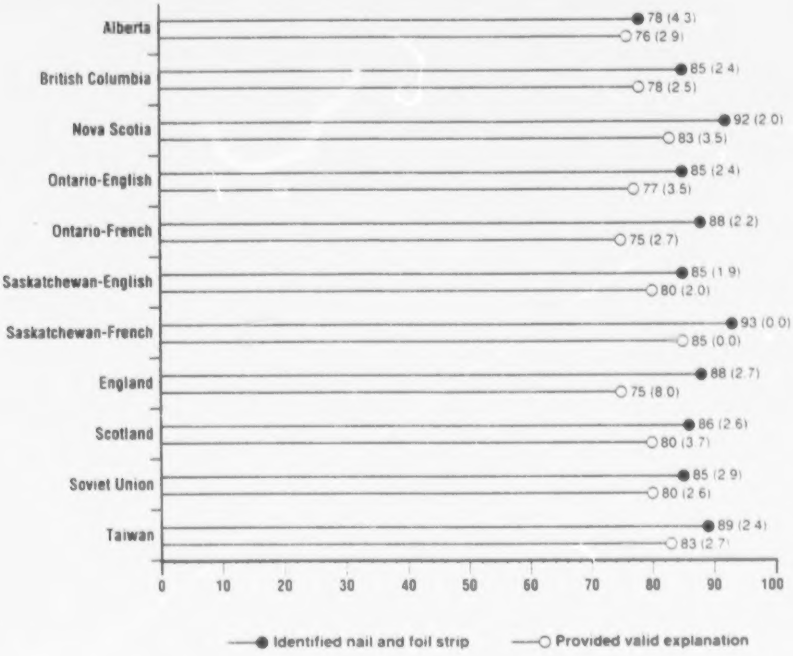
Problems

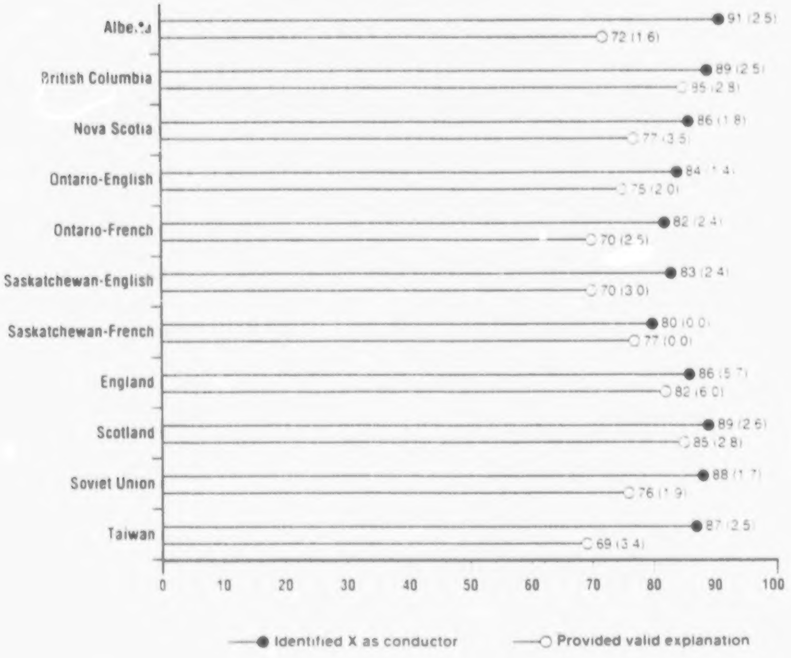
There was a problem in some Canadian provinces where the word "enable" in the instructions was read as "unable." Students who listed the nonconductors and provided an appropriate explanation were counted as giving the correct answers.

Comments

- Most students, 78 to 93 percent, categorized the objects correctly, but somewhat fewer were able to give a valid explanation for what they had done.
- In four of the countries and provinces, more students recognized the conductivity of object X than had categorized the original objects correctly and in two of these countries and provinces (and three in total), more students gave a valid explanation for their decision.

Percentage of Correct Responses (with Standard Errors)





CIRCUIT

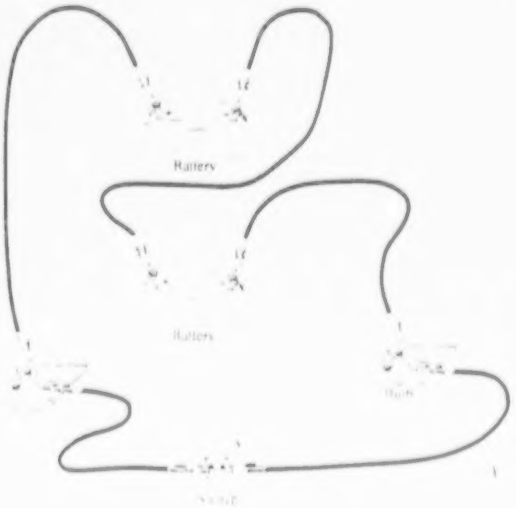
Task Descriptor

To construct an electrical circuit as represented in a drawing by selecting appropriate components and connecting them correctly.

Equipment/Material

Drawing of the circuit and a set of components as listed below. (Number of components required to construct the circuit are shown in parentheses.)

3 batteries (2)
2 battery holders (2)
3 bulbs (2)
2 bulb holders (2)
1 switch (1)
6 wires with clips (5)



Student Instructions

Use the objects on the card to make up the circuit shown in the drawing. You may not have to use all of the equipment. When your circuit matches the diagram, close the switch and see what happens. Raise your hand and ask the administrator to check your work.

Scoring Scheme

Credit was given for the correct positioning of batteries and bulbs, and for using five wires to form a closed loop, thus enabling the bulbs to light.

Problems

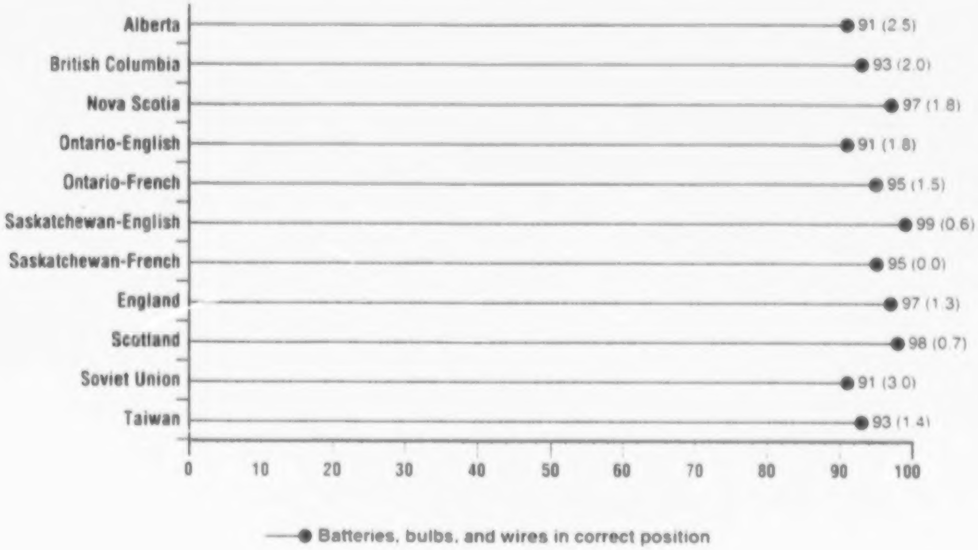
A loose connection in a bulb holder in one of the kits used in Ontario prevented the two bulbs from lighting, but students were credited for constructing the circuit correctly.

Comments

- Almost all students across participating countries and provinces completed this task successfully.

BEST COPY AVAILABLE

Percentage of Correct Responses (with Standard Errors)



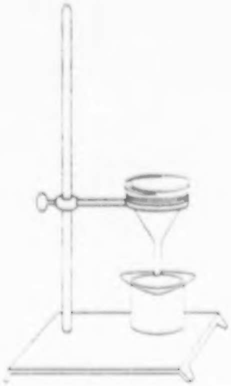
FILTER

Task Descriptor

To set up apparatus for filtering, as shown in a drawing, and to filter some muddy water.

Equipment/Material

A ring stand, a funnel, a beaker, and a folded filter paper. Also, a bottle of muddy water.



Student Instructions

Set up the apparatus as shown in the drawing above, put the folded filter paper into the funnel, and pour a small amount of muddy water into the funnel. Raise your hand when you have gotten some clear water and ask the administrator to check your work.

Scoring Scheme

Credit was given for the apparatus being assembled correctly, the filter paper being inserted correctly in the funnel, and for any clean water obtained.

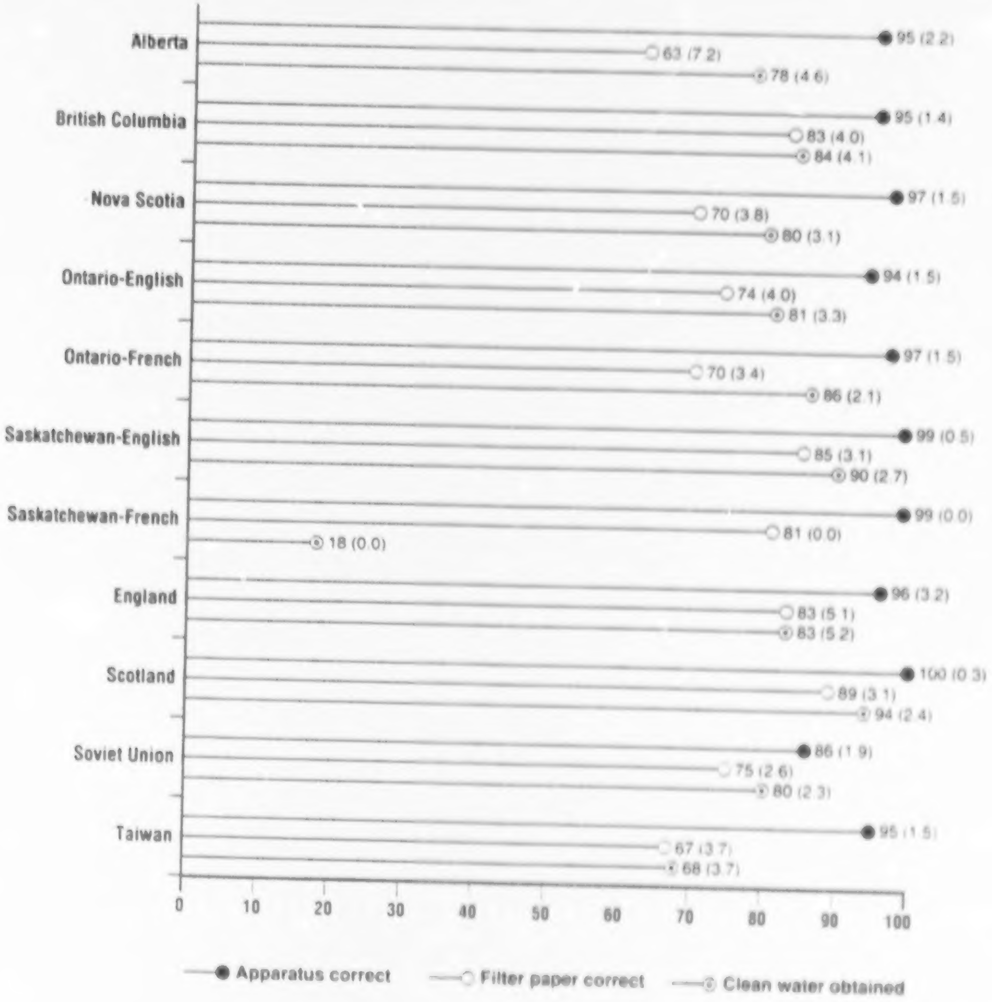
Problems

In the pilot-testing, filter papers were supplied unfolded and this caused widespread problems, but in the final assessment they were pre-folded.

Comments

- There was a high success rate, 86 to 100 percent correct, in assembling the apparatus; but more difficulty was experienced with correctly inserting the filter paper, where success ranged from 63 to 89 percent correct.
- Despite problems with the filter paper, many students were still able to obtain some clean water.

Percentage of Correct Responses (with Standard Errors)



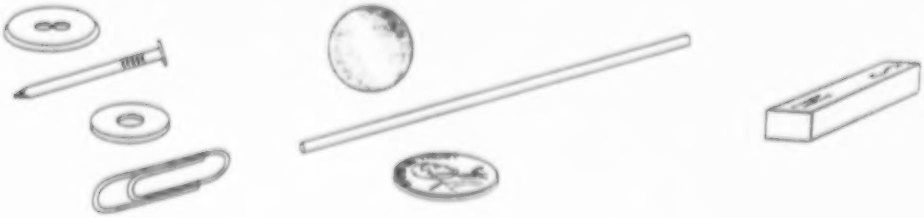
MAGNET

Task Descriptor

To use a magnet to identify magnetic and non-magnetic items and then to explain the difference between them.

Equipment/Material

A magnet and the following seven objects: plastic button, iron or steel washer, steel paper clip, iron nail, glass marble, plastic rod and copper coin.



Student Instructions

Test the objects with the magnet and divide them into two groups. List the objects in the two groups and explain what makes the objects in the two groups different.

Scoring Scheme

Credit was given for grouping the objects correctly. Four categories of explanations were recorded: namely, that one group was made of iron or steel, that one group was attracted by the magnet, that one group was made of iron and steel and was attracted by the magnet, and any other explanation.

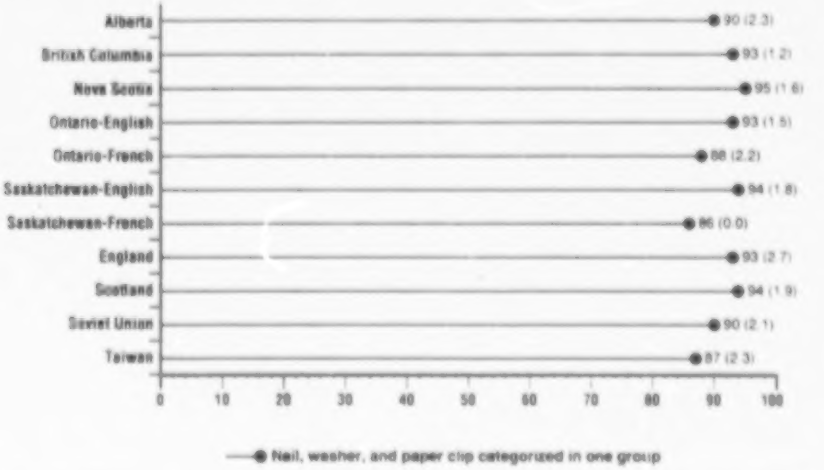
Comments

- Generally students performed the categorization task well, scores ranging from 86 to 95 percent correct; but 10 percent of the students across all countries and provinces gave irrelevant explanations.

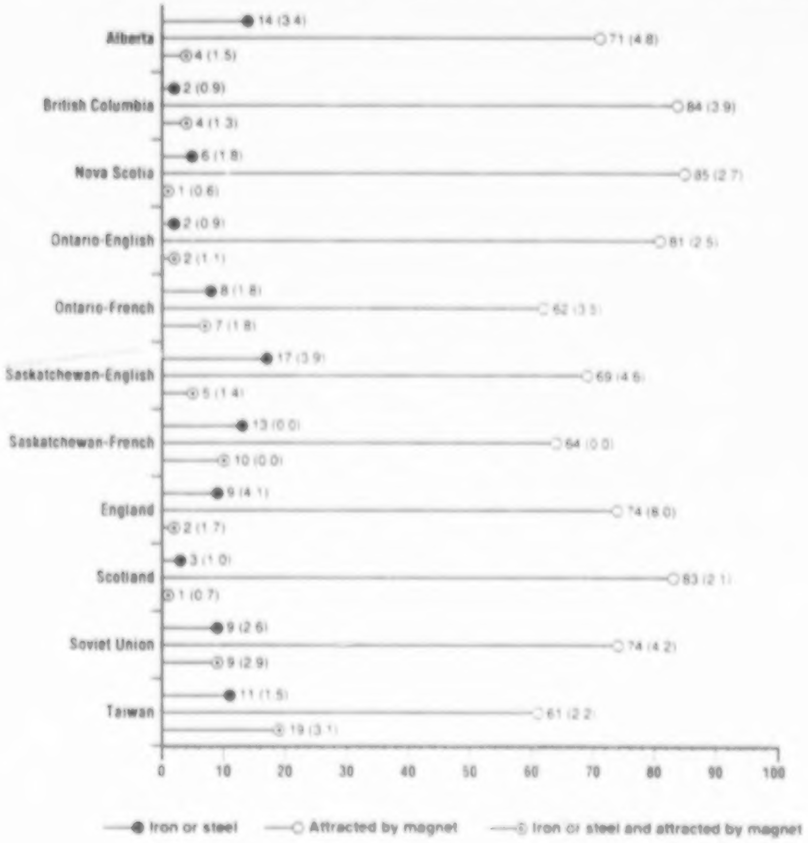
- Omission rates were generally low, but there was a 6 percent omission rate in England.

- The most frequent explanation for students' categorization was that one group of objects was attracted by the magnet; 79 percent of the students across participating countries and provinces gave this response. Fewer, between 4 and 30 percent, mentioned iron or steel, and this varied considerably among countries and provinces.

Percentage of Correct Responses (with Standard Errors)



Percentage of Students Giving Particular Explanations (with Standard Errors)



INDICATORS

Task Descriptor

To determine whether three solutions contain glucose, starch, or glucose and starch using indicators for glucose (test strip) and starch (iodine solution).

Equipment/Material

Three dishes labelled A, B and C containing the standardized, unknown solutions. Glucose test strips and iodine solution in a dropper bottle.

Student Instructions

The glucose test strip will turn from yellow to green on contact with a solution containing glucose and the iodine solution will turn blue-black when starch is present. The dishes A, B and C contain three different solutions which you are to test for glucose and starch using the indicators. Take the dish filled with solution A and dip the glucose test strip into it. Let the test strip dry. Add a drop of iodine solution to dish A. Observe all the results, report what solution A contains and repeat for solutions B and C.

Scoring Scheme

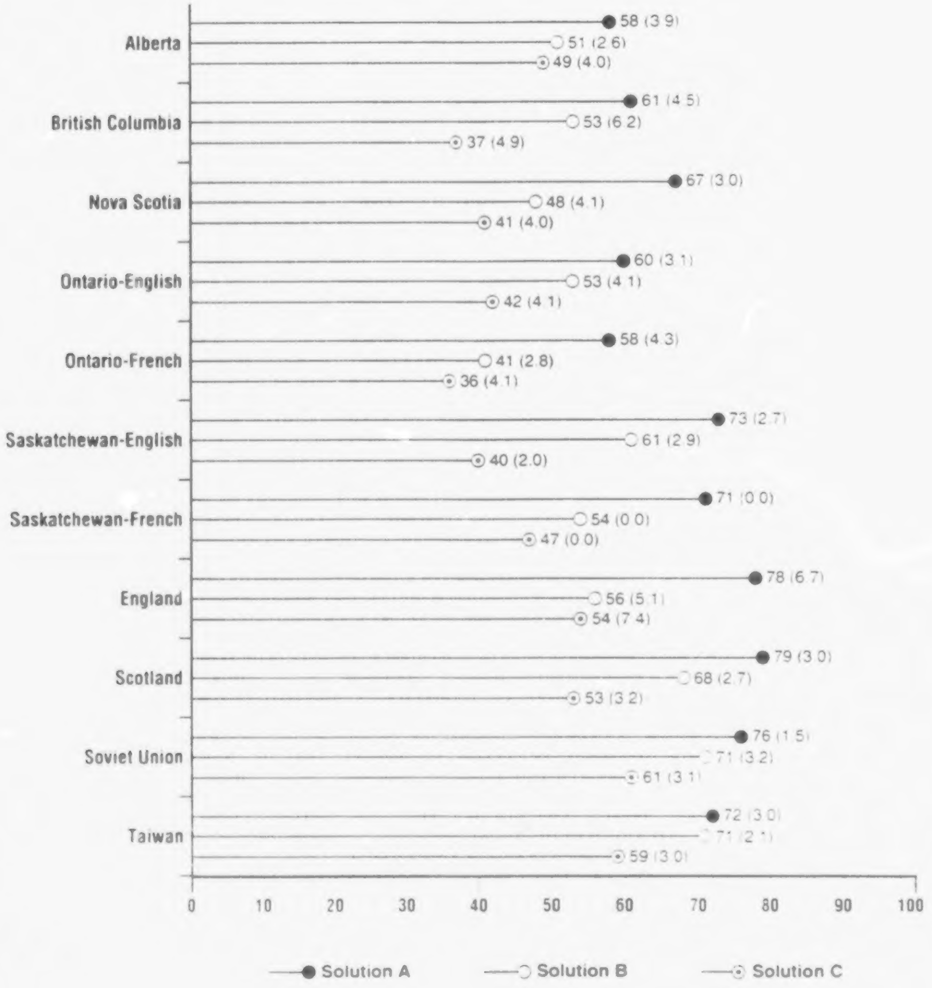
Credit was given for identifying glucose only in solution A, starch only in solution B, and glucose and starch in solution C.

Comments

- The differences in performance among countries and provinces were substantial in all three tasks. For each task, the difference in the scores of the highest and lowest performing populations was at least 20 points.

- Success rates in identifying the solution containing only glucose were highest, averaging 68 percent correct across participating countries and provinces. Those for the starch-only solution and the mixture of both averaged 53 and 47 percents correct, respectively.

Percentage of Correct Responses (with Standard Errors)



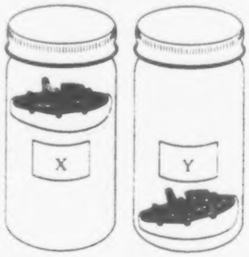
FLOAT

Task Descriptor

To select correct observations about flotation from two sets of objects.

Equipment/Material

Two small glass jars labelled X and Y containing clear liquids and identical plastic toys, one floating (in jar X) and one submerged (in jar Y).



Student Instructions

Look carefully at the two jars -- you may pick them up. Five other students looked at these jars and made the following statements. Which statements are observations, that is, they describe what the student actually saw?

A. I see a toy floating in jar X.
B. I see a toy floating in jar Y.
C. I see a toy in jar X that is made of a different plastic than the toy in jar Y.
D. I see jars containing colourless liquids and coloured toys.
E. I see a toy in jar Y that is heavier than the toy in jar X.

Scoring Scheme

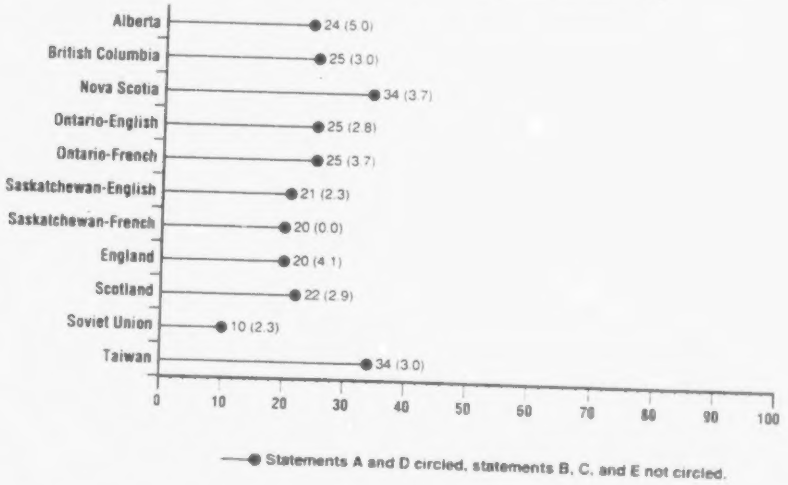
Credit was given for circling correct statements A and D and not circling incorrect statements B, C and E.

Comments

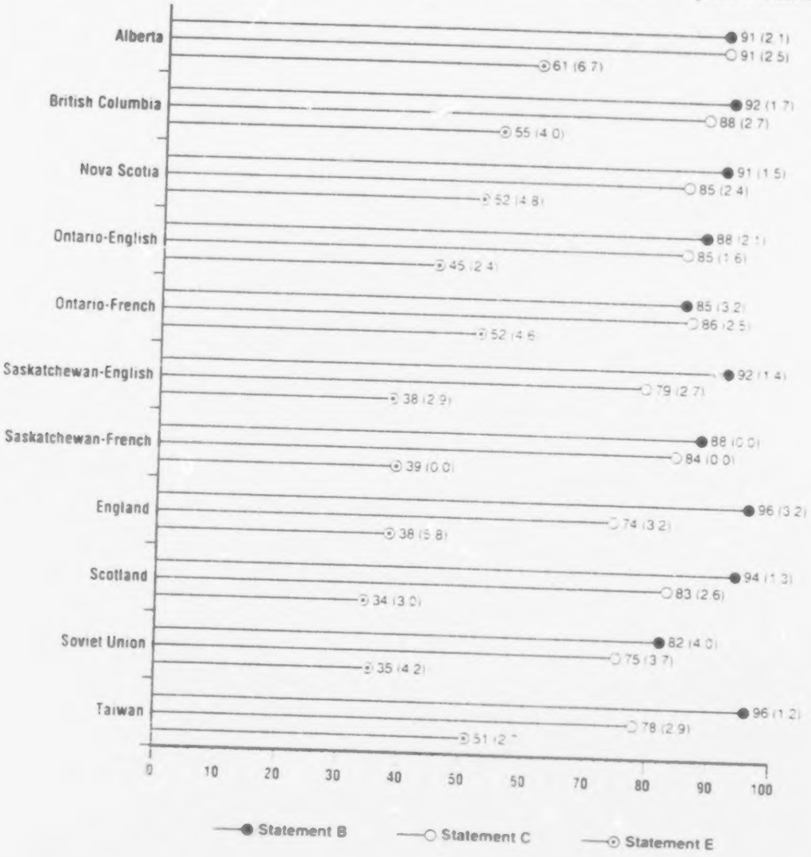
- The percentages of students who circled both correct statements *and* none of the incorrect ones were low, ranging from 10 to 34 percent.

- Most students recognized statements A and D as observations. Almost all students recognized that statement B, the opposite of statement A, is not an observation. Most students recognized statement C, that the two toys were made of different plastic, is an incorrect statement, probably because the toys looked so similar. However, statement E, that the mass of the toys were different, proved attractive to many students and they circled it, even though they had no way of knowing the mass of the two toys.

Percentage of Correct Responses (with Standard Errors)



Percentage of Students Recognizing Incorrect Statements (with Standard Errors)



TABLET IN WATER

Task Descriptor

To observe and record all the changes which take place when a tablet dissolves in water.

Equipment/Material

Water supply, plastic cup, and fruit-flavoured, coloured fizzy tablets.

Student Instructions

Observe what happens when the tablet is in the water. Write as many different things as you notice.

Scoring Scheme

Credit given for all appropriate visual, auditory, and olfactory changes recorded.

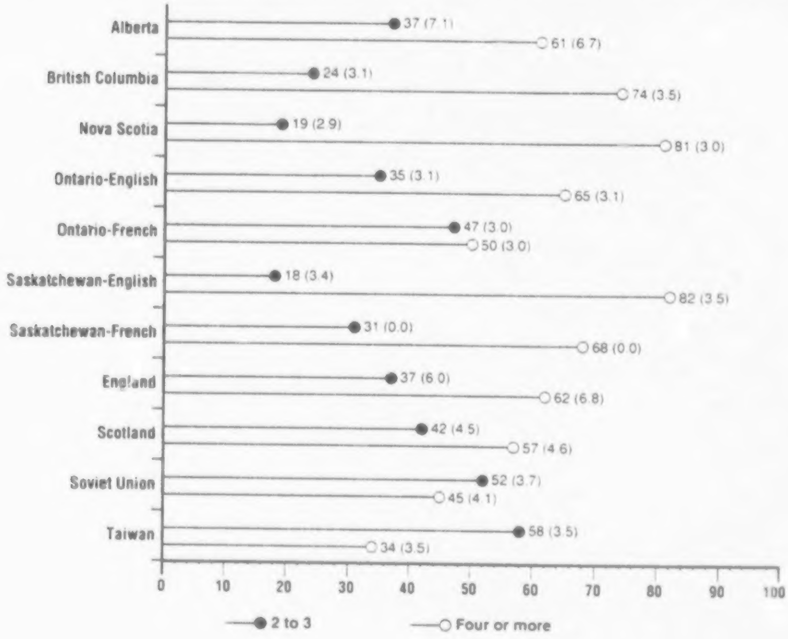
Comments

- The changes that were recorded by most students were in the size of the tablet, the colour of the water, and the bubbling of gas. These are all visual changes and it may be that the use of the word "observe" in the students' instructions biassed their responses towards such changes. However, there were substantial differences in the reporting of different visual changes and among different countries and provinces.

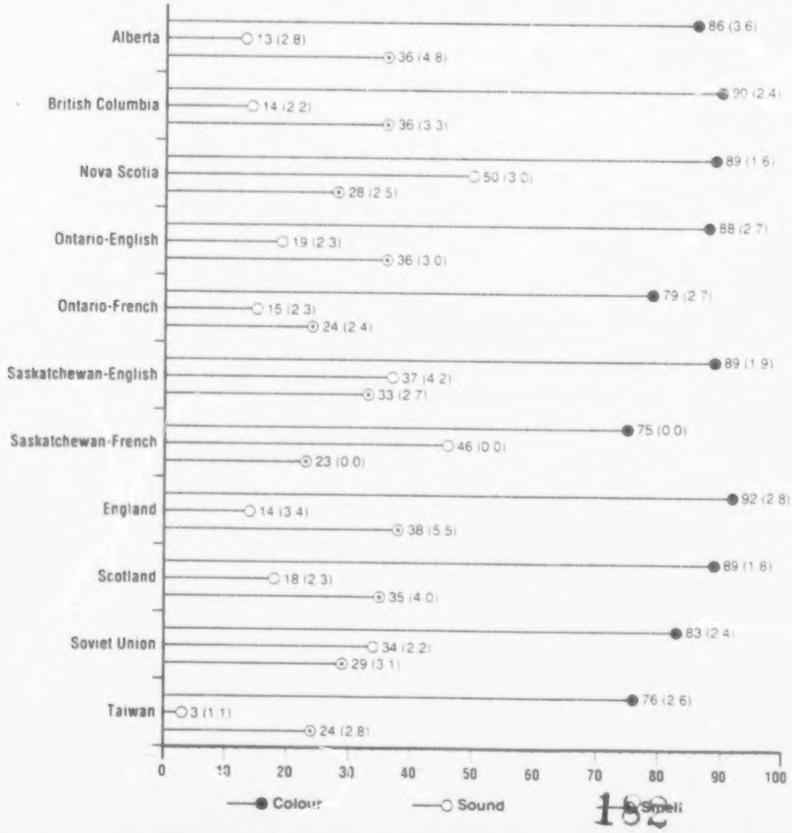
- A notable feature was a wide range in the reporting of the fizzing sound as the tablet dissolved, from 3 percent in Taiwan to 50 percent in Nova Scotia.

- At least one-half of the students in participating countries and provinces mentioned four or more observations, except in the Soviet Union and Taiwan, where the percentages were 45 percent and 34 percent, respectively.

Percentage of Students Mentioning Correct Observations (with Standard Errors)



Percentage of Students Mentioning Specific Observations (with Standard Errors)



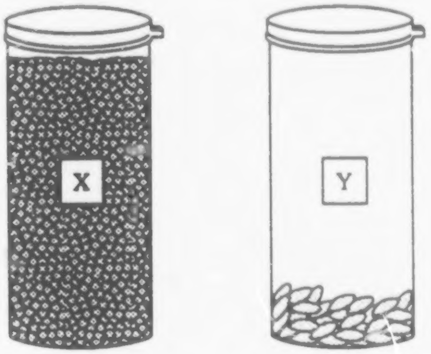
SEEDS

Task Descriptor

To categorize two different types of seeds according to their size, shape and colour.

Equipment/Material

Three groups of seeds labelled 1, 2, and 3 and two containers labelled **X** and **Y** with the "unknown" seeds.



Student Instructions

Your task is to decide in which group seeds X and Y belong and to state your reasons. Look carefully at the three groups of seeds 1, 2, and 3 -- you may pick up the containers.

Scoring Scheme

Credit was given for relating seeds X to group 3 and mentioning colour and size. Also for relating seeds Y to group 2 and mentioning shape.

Problems

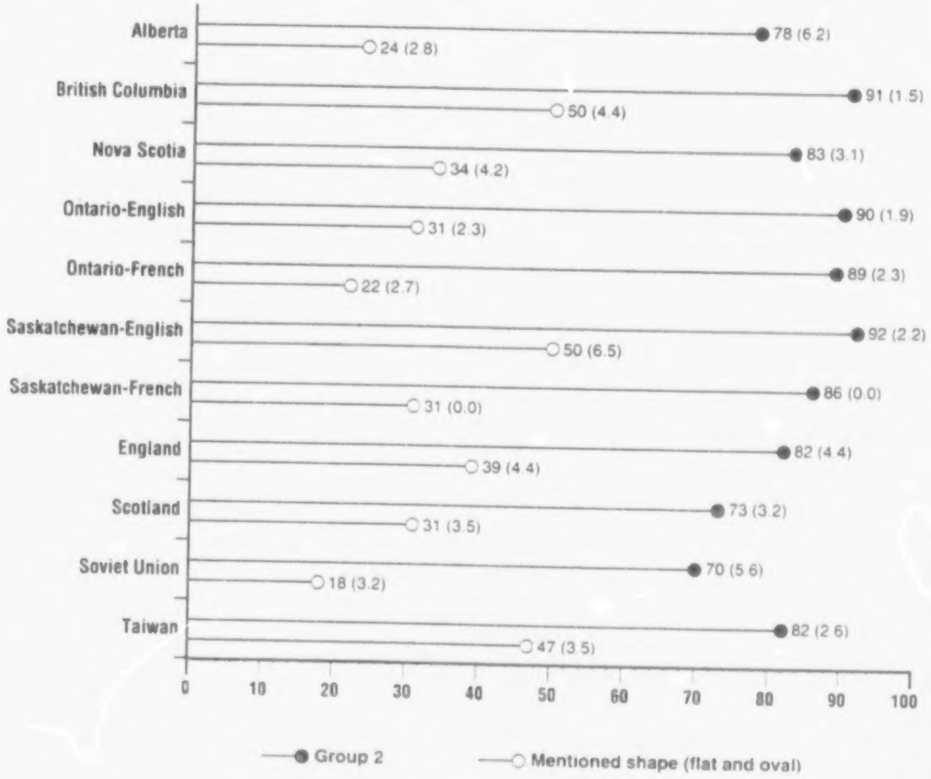
Seeds could not be included in the kits and sometimes it proved impossible to obtain comparable seeds in all of the countries involved. Because the colour of the sesame seeds varied (sometimes white, sometimes yellow), comparable scores could not be obtained for the first part of the task, categorizing seed X.

Comments

- In general, high proportions of the students were able to assign seeds Y to the correct group, ranging from 70 percent in the Soviet Union to 92 percent in English-speaking Saskatchewan.

- Many fewer students provided the correct reason for their categorization. The percentages doing so ranged from 18 to 50 percent.

Percentage of Correct Responses for Container Y (with Standard Errors)



"Measuring What's Worth Learning"
and
"Mystery Graphs"

Measuring Up: Prototypes for Mathematics Assessment

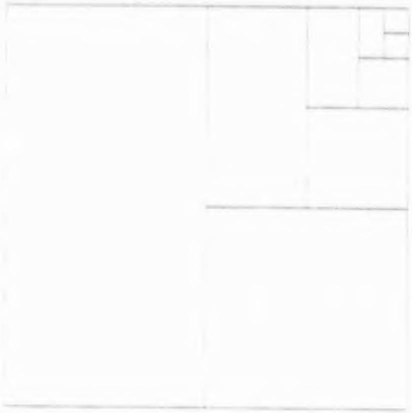
At the National Summit on Mathematics Assessment held at the National Academy of Sciences in 1991, Governor Roy Romer challenged the mathematics community to show through realistic examples just what we mean by "standards-based" education. *Measuring Up* contains 13 assessment prototypes that exemplify changes called for by the National Council of Teachers of Mathematics (NCTM) Curriculum and Evaluation Standards for School Mathematics. Two sections are reproduced here.

Copies of *Measuring Up: Prototypes for Mathematics Assessment*, Mathematical Sciences Education Board, National Research Council, Washington, DC: National Academy Press, 1993, may be purchased from:

National Academy Press
2101 Constitution Avenue, NW
Washington, DC 20418

Adapted with permission from *Measuring Up: Prototypes for Mathematics Assessment*. Copyright 1993 by the National Academy of Sciences. Courtesy of the National Academy Press, Washington, DC.

Perspectives on School
Mathematics



Measuring Up

Prototypes for Mathematics Assessment

Mathematical Sciences Education Board
National Research Council

NATIONAL ACADEMY PRESS
Washington, DC 1993

Measuring What's Worth Learning

The spotlight of educational reform continues to sweep across the stage of mathematics. First curriculum, then teaching, and now assessment have come under intense professional and public scrutiny. Amid deteriorating public confidence in the quality of American education, the mathematical community is addressing multiple challenges to articulate and implement effective standards in the key arena of testing, assessment, and accountability.

In the center of the assessment stage are three elements contesting for leadership. Conventional testing offers comfortable short-response tests on traditional content that are taken by millions of students every year. Reformers, including authors of the two K-12 *Standards* documents from the National Council of Teachers of Mathematics (NCTM), call for fundamental change — different in content, in format, and particularly in spirit. To this well-rehearsed contest of traditionalist vs. reformist has now been added a third movement arriving from outside the educational community: the political call for assessment of progress towards our nation's new standards in mathematics education.

In the decade since publication of *A Nation at Risk*, the United States has moved a long way toward a new consensus for education. Talk of national standards, once taboo, is now commonplace; so too is talk of alternative school structures

BEST COPY AVAILABLE

and innovative licensure for teachers. It is now time to develop a new national understanding of standards-based performance as the true measure of educational progress.

Throughout this decade, mathematics has led the way in educational reform. The 1989 MSEB publication *Everybody Counts* was followed in just two months by publication of the NCTM *Curriculum and Evaluation Standards for School Mathematics,* with its theme of developing mathematical power in all students. Undergirding these reports are three fundamental principles of testing, assessment, and accountability:

- Tests should measure what's worth learning, not just what's easy to measure.
- Progress depends on constant correction based on feedback from assessment.
- Schools are accountable, both to taxpayers and to students.

Even as the renewed public scrutiny compels educators to demonstrate that children are learning, the NCTM's *Standards* require new ways of measuring what is being learned. Because the linkage between tests and teaching is so close, it is vitally important for the United States that assessment be based on instruments that are properly aligned with the goals of the *Standards.*

The Challenge

At the National Summit on Mathematics Assessment in April 1991, Governor Roy Romer, in his capacity as Co-chair of the National Education Goals Panel, challenged the mathematical community to show the nation what mathematics educators mean by mathematical power and what new and more demanding standards will mean for our young people. One month later, the MSEB authorized creation of *prototypes* of tasks that could be used to assess fourth-graders' mathematical skills and knowledge, thereby providing examples of what children educated according to the new standards should be able to do. They wanted to be sure that the voice of mathematics was heard early and clearly in the assessment reform movement. The MSEB determined that it should be prepared to show, by

Why we are doing this

example, the type of assessment exercises that would be appropriate to measure our nation's progress toward the goals of mathematics education.

To create the prototypes, the MSEB subsequently convened a small writing group of mathematics educators, teachers, and mathematicians. Taking up Governor Romer's challenge, the writing group created a sampler of tasks to encompass many of the goals for mathematics instruction that are expressed in the NCTM *Standards.* These tasks, which illustrate what a standards-based education really means, have been pilot tested on a limited basis in four states. Many have been revised, often more than once, but all can benefit from continued improvement and adaptations.

Readers who skip ahead will see that these prototypes are not only innovative and challenging but also just plain fun. Teachers, children, and even parents should find these tasks both engaging and surprising. We invite readers to try them, either before or after reading the surrounding analysis.

The Criteria

Not surprisingly, the MSEB writing group debated extensively the criteria for prototypical assessment tasks. They faced the pioneer's challenge — to use incomplete information as a basis for decisions whose consequences are difficult to foresee. From these discussions emerged several criteria that helped shape the nature and selection of prototypes in this volume:

What we are trying to do

- *Mathematical content:* The tasks should reflect the "spirit" of the reform movement, but not necessarily be limited by particular curricular content, present or planned. Many of the tasks should incorporate a variety of mathematics, particularly in areas such as statistics, geometry, and probability that are least often emphasized in traditional K-4 programs.
- *Mathematical connections:* Everyone involved in the mathematics reform movement, from classroom teach-

ers to national policy makers, agrees on the importance of connecting mathematics — to science, to social science, to art, to everyday life, and to other parts of mathematics. Accordingly, the prototypes should develop links with science, with the visual arts, and with the language arts.

- *Thoughtful approaches:* Insofar as possible, the tasks should promote "higher-order" thinking. Just as the verbs explore, justify, represent, solve, construct, discuss, use, investigate, describe, develop, and predict are used in the *Standards* to convey "active physical and mental involvement of children" in learning mathematics, they are appropriate to seek in assessment activities as well. Further, given a choice between cognitive complexity and simplicity, the focus of these tasks should be on the former.
- *Mathematical communication:* The tasks should emphasize the importance of communicating results — not simply isolated answers, but mathematical representations and chains of reasoning. Children should have opportunities to work in groups to explain their thinking to others, and to write explanations of their approaches.
- *Rich opportunities:* The tasks should let children solve problems via a variety of creative strategies; demonstrate talents (artistic, spatial, verbal) beyond those normally associated with numerical mathematics; invent mathematics that (to them) is new; recognize opportunities to use and apply mathematics; and show what they can do (as opposed to what they cannot do).
- *Improved instruction:* The tasks should have the potential for influencing instruction positively, both in content and in pedagogy. Teachers who use these tasks should become better teachers as a result of the experience; children who participate should emerge with increased self-confidence and heightened expectations for future mathematics courses.

The Caveats

These tasks are *prototypes*, not tasks ready for immediate administration to fourth-grade students. They are intended to illustrate possible directions for new assessment instruments, not to be an example of a real assessment. Certainly they should be viewed as work in progress, not as fully completed blueprints.

Criteria related to cost, efficiency, and immediate feasibility were deliberately not imposed on the work of the writing group. These are important considerations for implementation, but not for this volume. The MSEB goal for *Measuring Up* is to promote long-term change, not to write assessment material for current courses.

As assessment instruments, these prototypes are intended for children who have had the full benefit of a *Standards*-caliber mathematical education in kindergarten through fourth grade. Hence the tasks as presented here will be more appropriate, generally speaking, for students of some time in the future. From the perspective that has historically dominated U.S. testing, these prototypes illustrate directions for tomorrow, rather than tasks for immediate practical use. From a perspective more common in Europe — where tests, appropriately publicized in advance, set targets for teaching and learning — these prototypes do serve the immediate purpose of defining appropriate goals for fourth-grade instruction.

What we are not trying to do

Moreover, the prototypes, as a set, are not intended to illustrate a single assessment that treats all of the mathematics important at the fourth-grade level. Much that is important in the curriculum is not covered adequately in the particular examples chosen for this volume. Nevertheless, to expand our view of appropriate mathematics goals for the primary grades, these tasks provide more opportunities for children to demonstrate their ideas in areas often missing from the curriculum (e.g., data, geometry) than in areas already well entrenched (arithmetic). The imbalance in these examples reflects our desire to illustrate the new, not an effort to reshape the curriculum to fit this particular set of examples.

These prototypes, which are tasks to be done in time spans ranging from one to three class periods, represent only one of many important forms of assessment. Other forms of assessment are essential for a balanced program, including *projects* (extended pieces of mathematical investigation designed to take a substantial block of time), *portfolios* (structured collections of student work gathered over a long time period), and *tests* (time-limited responses to shorter tasks). Some of the references at the end of this volume (e.g., Pandey [1991]; Stenmark [1989]) describe these alternative approaches.

The Audience

Whom we are trying to reach

Many readers of *Measuring Up* will be persons who are professionally concerned with mathematics education, particularly developers of tests and other assessment instruments. For such people, both those who work within commercial test development companies as well as those in educational settings at the state or local levels, *Measuring Up* should stimulate development of new approaches to assessment that reflect the broad goals of the nation's standards for mathematics education.

If mandated assessments evolve to resemble more closely the ones suggested in this book, it is clear that different approaches to instruction and testing will be needed. Hence school administrators and educational policy makers will also be affected by the changes implicit in these prototypes. The tasks will convey to the audience of policy makers and education leaders what mathematics educators mean by assessment reform.

A third audience for *Measuring Up* consists of classroom teachers, and not just those at the fourth-grade level. It is only natural that many practicing elementary school teachers may find some of these tasks to be somewhat daunting, especially if their students have not had the mathematical preparation that the tasks assume. Teachers should look at the prototypes not as current expectations, but rather as goals to aim for. The prototypes can be viewed both as examples of what tomorrow's assessment in mathematics might be like, and as examples of what today's goals for instruction should be like. In the meantime, teachers can use them as ideas for instructional activities for today. (A list of resources for teachers including the names and addresses of contacts in each state appears at the end of the volume.)

Another audience is the community of university-based educators who are responsible for the pre-service education of prospective teachers. They will find *Measuring Up* to be a source of ideas to use today for connecting the tenets of the mathematics education reform movement to classroom practice.

Finally, of course, the ultimate audience for these new assessment tasks and the ideas that underlie them is the elementary school children for whom the tasks were designed. The tasks provide good examples of challenging mathematical problems and situations that effective teachers can use even now as part of their instructional strategies. Today's children can begin to see the challenge in authentic mathematical problems even before tomorrow's tests give them an opportunity to demonstrate their accomplishments.

The Prototypes

Measuring Up contains thirteen assessment prototypes that exemplify changes called for in the *Standards*. In some cases the particular settings or contexts for the tasks are original, while in other cases some aspect of the task has appeared in another form previously.

The tasks in *Measuring Up* are intended for a largely hypothetical audience: fourth-grade children who have had a K-4 mathematics experience fully consonant with the NCTM *Standards*. Unfortunately, very few U.S. fourth graders have had the benefit of such an education. This is, of course, the whole point of the reform effort. One would not expect many of today's fourth graders to do very well on these tasks. Nonetheless the aim was to keep the tasks *accessible* to most of today's fourth graders; they should at least be able to understand what the tasks are about and become engaged in working on them.

What we have accomplished

Too often test questions and assessment tasks are presented solely in written form, which may be a burden for poor readers and for children whose first language is not English. Such children might not be able to respond to the tasks in a way that shows their true level of mathematical knowledge or skills. Many alternative presentations are possible: videotaped introduction; teacher-taught introduction; computer-based presentation; and presentation using manipulative materials. The prototypes illustrate each of these alternative modes of presentation, and two of the tasks are written in Spanish as well as in English.

Notwithstanding the possible variety in presentation, the prototypes in *Measuring Up* adhere to a certain uniformity of structure. Most are organized as a sequence of questions, often of increasing difficulty. On the one hand, this provides a structure around which the child's problem solving must be organized. On the other hand, this sequence of questions may suggest that the problem-poser, rather than the problem-solver, is in charge of the problem-solving process. Although other forms of organization are certainly possible, these prototypes provide sufficient imposed structure to help the mathematically less sophisticated student get started and show what he or she can do, while allowing plenty of open-ended space at the top to challenge the more advanced student. Even though the questions within a task often grow in difficulty, many of the tasks involve problem solving, reasoning, and communication right from the beginning. These are important aspects of mathematics for all children.

Just as the tasks are presented in several formats, so they are also designed to give children a chance to respond in a variety of modes — perhaps by constructing an object or by creating a pattern on a computer screen. One important response mode that is not specifically included in these prototypes is that of the child talking individually to a teacher, explaining his or her solutions orally rather than in written form. Pilot testing of the tasks has shown that children who have not had considerable experience in organizing their thoughts on paper find it much easier to tell someone else what they are doing than it is to record it in writing. Teachers who use tasks like the ones in this collection for their own informal assessment of how children are progressing mathematically will want to supplement written responses with spoken ones. In fact, asking a child to explain a solution in two forms — spoken and written — can help the child to sharpen and deepen both responses.

These prototypes can be used either for informal classroom-based assessment by an individual teacher, or for more formal external assessment, although certain modifications may be necessary to make the tasks suitable for a given purpose. All of the prototypes call for responses that go well beyond simple numerical answers, and most require the student to explain underlying patterns, relationships, or reasoning. As a result, the same activities can be useful to an individual teacher as she or he tries to discern more deeply how students are progressing mathematically, and to a district to discern the effectiveness of its instruction.

As the NCTM *Standards* urge, assessment should be embedded in instruction, so that most children would not recognize the assessment activity as a "test." Even when certain tasks are used as part of a formal, external assessment, there should be some kind of instructional follow-up. As a routine part of classroom discourse, interesting problems should be revisited, extended, and generalized, whatever their original sources.

Increasingly, educators are recognizing the value of having children work together in groups. Certainly group work exemplifies the NCTM's goal of stressing mathematics as a means of communication. Some of the tasks in *Measuring Up* are designed to be carried out in small groups, while in other cases, small groups are certainly a reasonable option. Continuing experimentation will be required to determine how the children can best be grouped for assessment tasks like these, and how to weigh individual vs. group work in performance evaluation.

In several cases the problems suggested here for fourth grade could also be asked in the eighth or even the twelfth grade, although naturally the expected sophistication and completeness of the responses would be very different. If a mathematical task is genuinely interesting and worthwhile for fourth graders, there is no reason why it should not be worthwhile for older children, or even for adults.

The Tryouts

What we learned from children

Each prototype was tested on several score fourth-grade students in a number of different locales. These "tryouts" were not designed to be either representative or comprehensive, but to aid in improving the tasks. This they did, but they did much more as well. By observing how students react to the prototypes, we learned much about the gulf that separates current students from the goals of the *Standards*. We also learned that we are novices on how these new forms of assessment will work in the classroom.

Three examples can illustrate the types of surprises that all teachers will encounter as they begin to explore and extend these prototypes:

- In a few cases the tasks as originally presented seemed not to be sufficiently challenging. One example is the "Lightning" task in which a fairly large proportion of the students could easily handle the map-reading requirements. So a question dealing with locating a lightning bolt that is a given distance from two observers was added.
- Sometimes a proposed task yielded no information of any interest at all. In "Bridges," there was originally a more open-ended question in which students were to create their own bridges. Nobody created anything that went even a little bit beyond the two-support, single-span examples. This may have been due to the wording of the question, to the backgrounds of the particular students, or to some other factor. This lack of inventiveness and perseverance is something worth pursuing since creativity is

an essential part of doing mathematics, for fourth graders as well as for everyone else. However, since the question produced virtually no information, it was dropped.

- One whole prototype was dropped entirely. It was a task on what is known as "Pick's Theorem" — which relates the area of a polygonal region on a geoboard to the number of nails on the boundary and in the interior of the region. The task was extremely open-ended and required extensive interaction between the teacher and individual students or small groups of students. Even if one assumed (as we do) that the teachers involved in the assessment are uniformly well versed in the subtleties of the underlying mathematics, there seemed to be no way of separating the effects of the teacher from the progress that individual students might make on the task. Perhaps such a task could be viewed as an assessment of the teacher-class unit, but in any case it seemed to be too problematic to include in this collection.

The Format

Each of the thirteen tasks is presented using the same outline. After the title, there is a suggested *time allotment*, which can vary from one to three class periods. This is followed by a suggested *student social organization*, although in many cases the task does not depend substantively on a particular form of grouping.

How we present the prototypes

Next comes the task itself. First there is a description of *assumed background*. In most cases this refers to specific aspects of the children's mathematical background, assuming — hypothetically, of course — that the children have had a K-4 education that fully meets the NCTM *Standards*. Second, there is a section on *presenting the task*, which details exactly what the teacher (or other assessor) should do. Finally, there is the *student assessment activity*. Very often this involves one or more sheets of paper on which students record their responses. (To reproduce these pages, which were scaled to the 7" x 10" page of this volume, the copying machine should be set to magnify them appropriately.)

The next major section is a *rationale* for the mathematics education community, which in many respects is the heart of *Measuring Up*. This is where comments on the content, style, or intent of the task appear (e.g., why the task was included), as well as more general messages about mathematics education that the task is intended to convey.

Following the main presentation of the rationale for the task, there are two subsections that provide further information. The first, task design considerations, discusses some of the details behind the task — why certain questions were phrased as they were, or why particular numbers were chosen. The second, variants and extensions, hints at other directions in which the task could be taken, for purposes either of instruction or further assessment. These subsections are far from exhaustive, for often the tasks could be starting points for weeks of instruction. One important message conveyed by this section is that these particular prototypes are in no way unique.

The next section describes a rough scoring system — what is called a *protorubric* — for the task. It is now widely recognized that an assessment task by itself means little without an indication of how children's responses would be scored. In other words, an important component of an assessment task is a scoring rubric that describes and orders a variety of answers that a child might typically give. For reasons discussed in the next section, the rubrics given here are necessarily tentative and incomplete — whence "protorubrics."

Finally, in some of the tasks there is a section containing *references* to relevant sources.

The Protorubrics

Although each task in this volume contains commentary about scoring based on student work, for a number of reasons we have not developed fully detailed scoring rubrics:

How might fourth graders do?

- The intended audience for these tasks are students who have had a mathematical education that is different from what is commonly available in U.S. schools today.

Mystery Graphs

Explore sets rather than individual pieces of data

Broaden the view of mathematics appropriate for the 4th grade

Apply mathematics to real-life experience

Suggested time allotment
Less than one class period

Student social organization
Students working alone

Task

Assumed background: This task assumes that the children have had extensive experience in dealing with sets of data, and, in particular, are familiar with interpreting data that are represented in line plots.

Presenting the task: The teacher should distribute the student materials and read enough of it to be sure that the children understand the task. It is also important to stress that the "classroom of fourth graders" is some other classroom — not theirs. In the pilot, it was necessary to clarify that "cavities" in question 1a refers to both filled and unfilled cavities.

Student assessment activity: See the following pages.

BEST COPY AVAILABLE

Name ______________________________ **Date** ____________

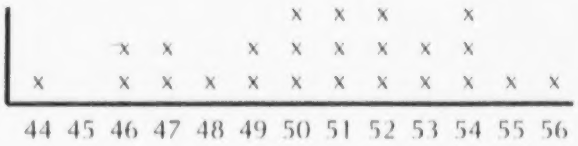
Look at the five graphs on the next pages. Each graph shows something about a classroom of fourth graders.

1. Which of the five graphs do you think shows:
   - **a.** The number of cavities that the fourth graders have? ______
   - **b.** The ages of the fourth graders' mothers? ______
   - **c.** The heights of the fourth graders, in inches? ______
   - **d.** The number of people in the fourth graders' families? ______

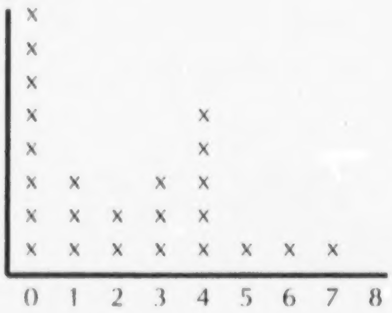
2. Explain why you think the graph you picked for **c** is the one that shows the heights of fourth graders.

3. Why do you think the other graphs don't show the fourth graders' heights?

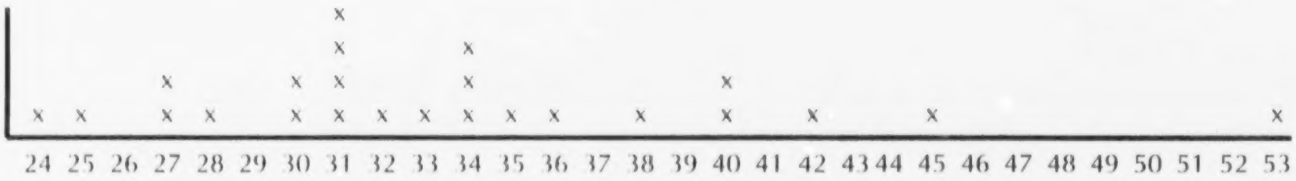
Graph 3



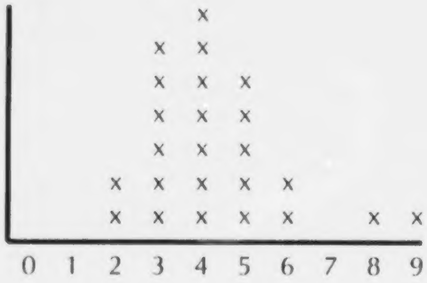
Graph 4



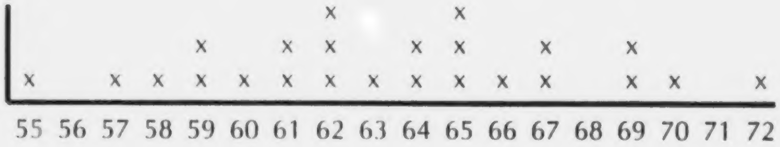
Graph 5



Graph 1



Graph 2



Rationale for the mathematics education community

This task puts a premium on looking at data sets, as opposed to individual pieces of information. This is a fundamental notion that should take an increasing role in the elementary mathematics curriculum. The task also gives children the opportunity to relate the graphical representations to their own experiences as fourth graders.

Ordinarily, of course, one would want children to have plenty of chances to collect, display, and analyze their own data, as the NCTM *Standards* suggest. If the task is going to fit within a single class period, however, there is not enough time to create five graphs for comparison. As a result, this task uses data that have already been collected from some hypothetical fourth grade. Clearly other assessment tasks (like the Hog Game and Buttons tasks in this collection) must include the collection, display, and analysis of data.

Task design considerations: Children seem naturally interested in data about people, particularly people of their own ages; this is one reason for choosing a hypothetical fourth-grade class as the basis of these data. The children will naturally bring their *own* experiences with heights, ages, family size, and dental health with them to the task. When using such situations for assessment purposes, one must be careful to use values of the data to which all the students can relate equally well. There may be cultural variations in family sizes or in the ages of fourth-graders' mothers, for example. To take this into account, the ranges of Graphs 1 and 5 are large enough to encompass every student's own family size and mother's age.

Questions similar to the one about heights could be asked about mothers' ages, family sizes, or cavities. The only reason such questions are not included is to save assessment time; the intent was to give an example of a task that could be done in less than one class period.

To some extent, this is a task that measures children's prior knowledge about the real world — about how many inches tall they are, how old their mothers are, and so on. If one is concerned with children's abilities to connect mathematics with their world of experience, this is a reasonable expectation.

The style of drawing line plots should be the same as the style to which the student is accustomed.

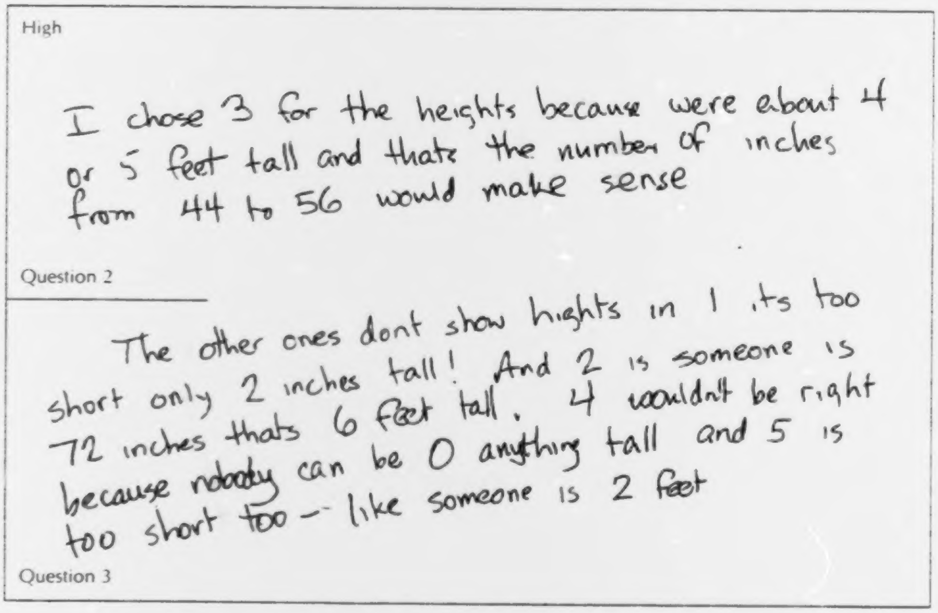
Ideally, the five graphs should be displayed so that the student can see them all at once.

Variants and extensions: A natural instructional follow-up to this task is to ask the students to compile data on heights, cavities, etc., from their own class, to compare with the data given.

Using just the data presented here, one could pose problems like: "Suppose Graph 2 really did show heights in inches. Whose heights could they be?" "Suppose Graph 3 showed the ages of the mothers of students in some grade level in our school. Which grade could that be?" "What other kinds of data could Graph 1 be showing?"

Protorubric

Characteristics of the high response:



High

I chose 3 for the heights because were about 4
or 5 feet tall and thats the number of inches
from 44 to 56 would make sense

Question 2

The other ones dont show hights in 1 its too
short only 2 inches tall! And 2 is someone is
72 inches thats 6 feet tall. 4 wouldn't be right
because nobody can be 0 anything tall and 5 is
too short too -- like someone is 2 feet

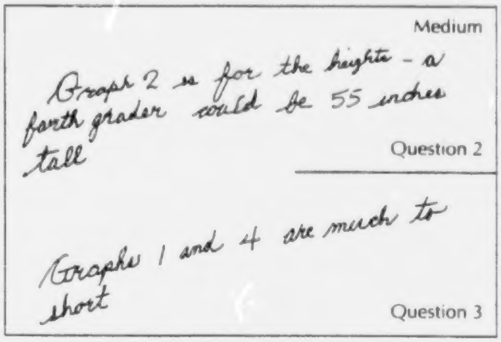
Question 3

The high response shows a full understanding of the relationship between the graphs and the data they represent.

The responses for question 1 are all correct (a. 4; b. 5; c. 3; d. 1). Questions 2 and 3, taken together, should explain that Graph 3 shows a reasonable range of fourth graders' heights, and that ranges of data in the other graphs are not as reasonable. The only real alternative candidate for the heights is Graph 2, but that would imply that there are fourth graders who are six feet tall.

Characteristics of the medium response:

Graph 1 and Graph 4 are interchanged (number of cavities and number of family members); or Graph 2 is used in place of Graph 3 or Graph 5; or Graphs 3 and 5 are interchanged. Nonetheless, graphs showing the correct general orders of magnitude are selected. Some portions of the student's justifications are reasonable.



Medium

Graph 2 is for the heights - a forth grader could be 55 inches tall

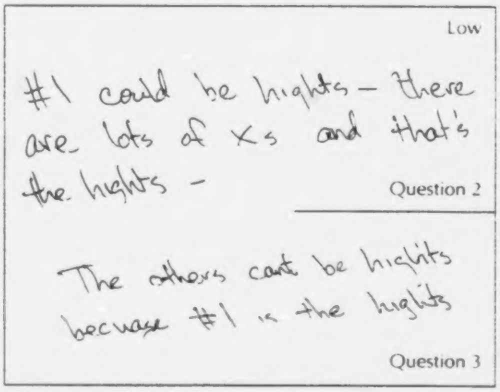
Question 2

Graphs 1 and 4 are much to short

Question 3

Characteristics of the low response:

At most one graph is chosen that shows totally unrealistic data (e.g., Graph 5, with a range from 24 to 53, is selected for the number of people in the families). Responses to questions 2 and 3 are missing or indicate that the student cannot interpret the graphs, or they do not show any reasonable sense of the magnitudes of more than one of the items.



Low

#1 could be hights - there are lots of Xs and that's the hights -

Question 2

The others cant be hights because #1 is the hights

Question 3

Reference

An earlier version of this task was developed by TERC (Cambridge, MA) for Education Development Center (Newton, MA).

"Piloting Pacesetter: Helping At-Risk Students Meet High Standards"

by Thomas W. Payzant and Dennie Palmer Wolf
Educational Leadership, February 1993

PACESETTER™, a new program of the College Board, is designed to raise expectations and improve performance of all American students. The program will provide secondary school course frameworks and related assessments in five subject areas, supported by professional development opportunities for teachers. All elements are being developed in cooperation with members of the leading national subject matter associations. The mathematics offering will be piloted in 1993-94, followed by offerings in English, world history, science, and Spanish.

The following article discusses PACESETTER™ as it is being pilot tested in the San Diego City Schools.

Payzant, Thomas W. and Dennie Palmer Wolf (February 1993). "Piloting Pacesetter: Helping At-Risk Students Meet High Standards," *Educational Leadership*, 50, 5:41-45. Reprinted with the permission of the Association for Supervision and Curriculum Development. Copyright © 1985 by ASCD. All Rights Reserved. Reprinted with the permission of Thomas W. Payzant.

Piloting Pacesetter: Helping At-Risk Students Meet High Standards

Thomas W. Payzant and Dennie Palmer Wolf

The San Diego City Schools, in partnership with the College Board, are piloting a program that seeks to prepare *all* students for the educational demands beyond high school.

[illegible]

Martin is 14. He reads on a 4th grade level. His writing is simple—not because he doesn't have complex thoughts—but because he often struggles to find the English word he wants, and 40 minutes simply isn't enough time to think, draft, and revise. He wants to graduate from high school and enter a demanding job-training program at a local light and power company. As his father points out, "It's the difference between $6 and $20 an hour all the rest of your life."

But the entry test is no joke. To pass, you need the modeling skills to notice patterns and predict possible difficulties down the line in the machinery. That entails working with Boyle's and Charles' laws and algebraic equations, and diagnosing sources of possible error. And it doesn't end there. The company is looking for employees who are able to interview suppliers and examine product information and forms written in Spanish, Japanese, or German.

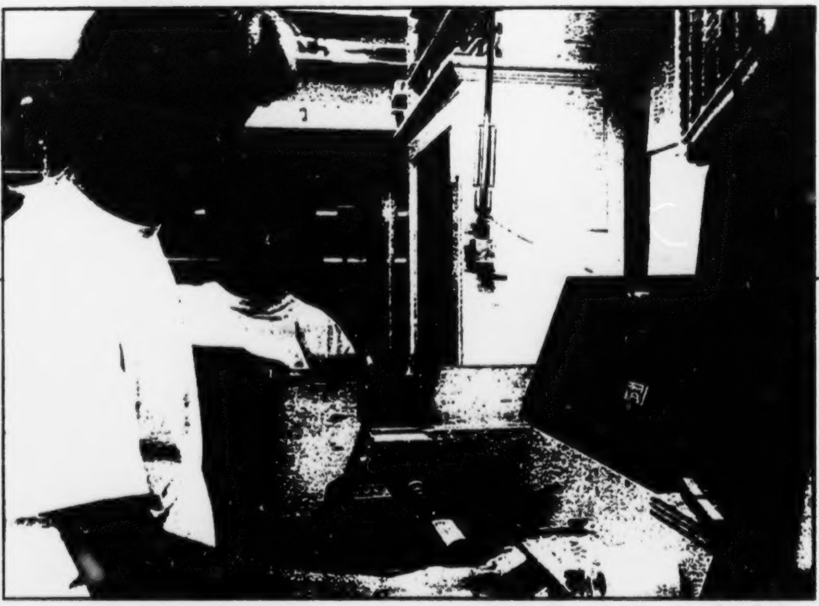
Access to High Outcomes

Gone are the days when graduation was a matter of going to school just enough to earn your Carnegie credits, or when any high school diploma could act as a passport. Public high schools, like those in San Diego, have as their major imperative helping *all* students prepare for postsecondary education—in colleges, in public service, or on the job, where the ticket is high-level competence, not attendance. The challenge is daunting. San Diego is an urban district of 125,000 students with diverse racial, ethnic, linguistic, and socioeconomic backgrounds. Sixty different first languages are spoken: 30 percent of the students are Hispanic, 19 percent are Asian (with large Indo-Chinese and Filipino groups), 16 percent are African American, 34 percent white, and 1 percent "other."

In this context, we have had to rethink traditional approaches to equity. We can no longer be content solely with the simple arithmetic of inputs—racially mixed schools, racially diverse teachers, classes of equal size, and bilingual opportunities for learning. We now face the challenge of providing equity of *outcomes*. This is a tall order in American public schools, where there is a long-held belief that ability is distributed in a normal curve pattern and, consequently, tracking is not only convenient, but appropriate. To uproot such deep beliefs demands a program of serious and sustained change in attitudes, daily practices, curriculum, and assessment.

In San Diego, we began five years ago by instituting a common core curriculum. Today, to be graduated from high school, a student must take four years of English, three years of math, two of science, three of social studies, and must meet a fine arts requirement. At the same time, we eliminated lower-level elective courses in English, math, and science. In mathematics, we established a pre-algebra/algebra sequence for all students, dropping all general, consumer, and business math courses.

As promising as these innovations are, by itself, this educational architecture won't promise Martin the life he and his family hope for. As a district, we have to guarantee more than coursework. We have to ensure that Martin encounters mathematics that is



[illegible]/San Diego City Schools

more than blind calculation and formula juggling. However, no urban district of our size and diversity has the dollars to guarantee these outcomes single-handedly. To provide excellence for *all* demands partnerships. We have to build on the standards the National Council of Teachers of Mathematics has developed, and we have to join hands with the social and natural sciences, as well as technology, to figure out the "big ideas" we ought to be concentrating on. But most critically, partners can help us think about the minute-by-minute invention of actual courses that can *enable* Martin—not merely remediate him.

A Push-Pull Strategy

If you say "College Board," most people think of an elite gate-keeping organization that decides who should go where with how much scholarship money. Not so. For the last decade, the College Board has been an active, vocal participant in school reform. Ten years ago, the board published *Academic Preparation for College* to inform students, teachers, and families about the necessary pathways to postsecondary education. In the ensuing years, the Educational EQuality Project (*E* for equality, *Q* for quality) developed workshops and publications to get the word out that more students deserved to attend, and could flourish in, college. In a second decade, the College Board has launched even bolder steps that add up to what has been called a "push-pull" strategy for major school reform. For example, the board, working with major educational foundations and a national consortium of researchers and teachers, has developed EQUITY 2000—a demanding program of pre-algebra, algebra, and geometry designed to ensure that minority students thrive in vigorous high school mathematics programs.

If EQUITY 2000 accounts for the "push" of this strategy, then the College Board's Pacesetter initiative accounts for the "pull." Through this program, the College Board is devoting major resources to determine how to make the high-standards curriculum, strong teaching, and performance assessment, long associated with its Advanced Placement Program, a part of every high school student's experience.

In San Diego, we have long used the AP Program as an equity tool. Unlike gifted and talented programs, these courses do not require cutoff scores or special certification: any willing student can enroll, and any teacher can take up the challenge of teaching a rigorous and inventive course. Characteristically, such courses focus on ideas and concepts and on helping students display their understanding in performance assessments (for example, applying physics principles to a novel situation and predicting possible outcomes). AP teachers often form professional groups, exchanging syllabi and teaching strategies and acting as readers when the open-ended portions of exams are graded. Not surprisingly, we have found these courses work toward equity, not elitism. They turn out to be laboratories for thinking through how excellent work might be demanded of a full range of our students.

Consequently, when the College Board proposed Pacesetter, we were more than interested. The project called for developing yearlong courses and associated assessments, along with detailed plans for teacher training, in mathematics, English, world history, science, and foreign language. Some courses would be keystones designed to integrate and deepen what students had learned throughout high school. For instance, in 12th grade science, students might conduct projects about complex issues that involved the merging of concepts and problems from earth science, biology, chemistry, and physics (for example, situations in which the chemical composition and the direction of flow affect how toxic waste takes its toll on the plant and animal life in a particular ecological niche). In 12th grade English, students might draw on their reading and insights from American, British, and world literature to trace the evolution of literature written in English from its origins to the present.

Other cornerstone courses, such as those in intermediate Spanish and world history, would suggest the kinds of knowledge and skills students should have midway through their high school careers. These worthwhile outcomes that addressed the chronic problem of differential access to knowledge would be worked on with national committees of skilled teachers, researchers, and members of national curriculum organizations. At the same time, as part of Pacesetter, we would be linked to six quite diverse pilot sites: Broward County, Florida; Prince George's County, Maryland; Battle Creek, Michigan; Charlotte-Mecklenburg, North Carolina; Irving, Texas; and Rutland, Vermont.

From Declaration to Realization

San Diego already has a history of innovation and a wealth of partners. Why take on more?

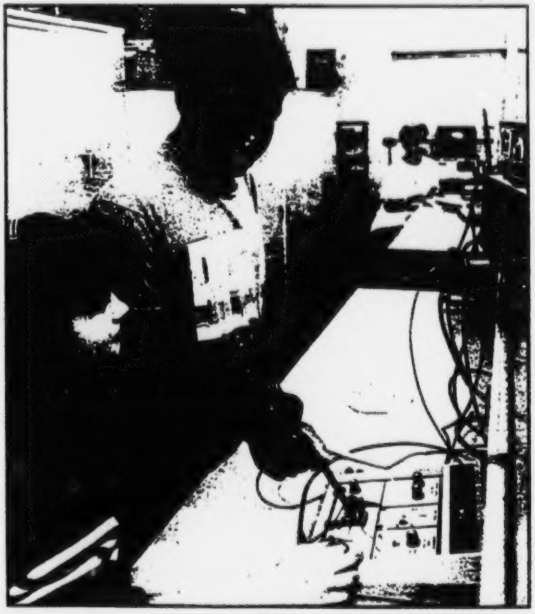
We are in the midst of a vigorous national effort to set standards. We have national educational goals for the year 2000. The National Council of Teachers of Mathematics has published widely regarded content standards. Social studies, foreign language, arts, and language arts teachers are headed in the same direction. Clearly, there is no shortage of statements about what we *ought* to do. What we lack is a clear, concrete vision of *how* to reach those goals. The issue for us as an urban district is not more declaration: it is realization.

Pacesetter is centrally about realization. At this moment, national committees of classroom teachers are designing specific course frameworks. English teachers are hotly debating how to give students entry to the major "cultural conversations" of our evolving culture. They are deliberating how to provide a background knowledge of writers like Shakespeare, without ignoring the fact that contemporary performances of *Othello*—set in Haiti or Los Angeles—could give new meaning to the play. Mathematicians are struggling to design a course that can offer pre-calculus students what they need and teach other students how to be critical consumers and become skillful at quantitative reasoning. World history teachers are grappling with how to use the concepts of climate, migration, and technology to make the study of history increasingly more global.

Each Pacesetter course will include:

1. an outline of subject content and anticipated learning outcomes developed by leading teachers and specialists from professional subject-matter associations and universities (for example, in the case of mathematics, the National Council of Teachers of Mathematics and the Mathematics Association of America);
2. teacher-training and support activities keyed to the content outline for each course, including in-school assessment techniques, summer institutes, workshops, and publications illustrating successful approaches to teaching diverse students;
3. classroom assessments that help teachers monitor and shape instruction while providing ongoing feedback to students;
4. end-of-course assessments (such as projects or portfolios);
5. a valid system for scoring end-of-course assessments on a state, regional, or local level.

But realization—even at this early stage—has to get beyond lists of ingredients to new visions of learning, collaboration with teachers, and assessment.



Barry Ladow/San Diego City Schools

Learning Outcomes for Students

Although the dust has hardly settled on the outcomes for Pacesetter English 12, early collaborations between the College Board and the National Council of Teachers of English are sketching a lively picture of what's to come. Students will read both classic and modern works in order to understand how we have framed and currently think about major human issues. Literacy, in this context, becomes not just the ability to decode and record, but to interpret and create a wide range of cultural texts—speeches, performances, written literature, documents, and even films.

At the outset of the course, students might introduce themselves, then play back what they have said about themselves and their lives—analyzing how words, images, and performances create specific impressions. Turning from their own oral expression, students will read short works from literature written in English, examining similar issues of self-presentation and representation through language. Moving on to larger works, students might read and watch productions of *The Tempest*, thinking about how self, familiar, and other (Prospero, Miranda, Ariel, and Caliban) are created through their own speech and what others say of them. Working in independent reading groups, students will investigate this legacy by looking at works as diverse as *Othello* or Toni Morrison's *Beloved*. Throughout the course, students will explore focal works that have shaped the way English speakers make sense of the world. British works as diverse as *The Tempest* and *Heart of Darkness*, American works that could range from early settlers' journals to *The Adventures of Huckleberry Finn*—as well as African, Caribbean, and Indian literature.

BEST COPY AVAILABLE

Throughout, students will take on the active roles of authors and critics, in addition to the familiar role of reader.

New Opportunities for Teachers

The 12th grade mathematics course focuses on what happens when we confront complex quantitative data sets with the need to understand patterns, continue research, or reach conclusions. In this setting, teachers' roles shift dramatically. They become researchers constructing rich "case studies" in which linear, exponential, and logarithmic functions can be applied to problems in fields like industrial design, economics, and demographics. For example, one member of the mathematics committee has proposed that students use mathematics to model the impact of major historical events. For instance, one problem might be "How different would contemporary Europe be if the Black Death had not occurred?"

Teachers are also designers, as they try these novel, more demanding approaches with students and assess how the materials work with a full range of students. What, for instance, does it take to get a student with a shaky mathematics background to apply reasoning capacities and questioning abilities he or she may have developed elsewhere?

Already by the summer of 1993, mathematics teachers from all seven sites will address the question of teachers' learning. Joining with teachers from the College Board's EQUITY 2000 project, they will examine what teachers need to know in order to become strong coaches and diagnosticians for students working in challenging mathematical environments. Subsequently, participants will assume the dual roles of instructor and critic, as they field-test a proposed sequence of applications that call for simple linear through complex logarithmic functions.

What is emerging from these efforts? A radically different view of professional development—no shrink-wrapped, teacher-proof materials to be swallowed whole the night before. If teachers are to become inventive users of the course frameworks and skilled assessors of student work, they must be actively involved in all stages of implementation.

New Questions About Assessment

Two conflicting purposes often crisscross assessment programs: the *responsibility* to use any assessment to respond to student work and encourage growth and the *demand* that assessment provide reliable, quantifiable information about student learning. As a nation, we have a long history of downplaying the first and highlighting the second. So consuming has our demand been for accountability data that we have often allowed rote and short-answer testing formats to obscure the potential richness of assessment. But if students like Martin are to realize their dreams, we need a more complex view of student assessment.

Pacesetter will allow urban districts like San Diego to take part in a broader national discussion about combining these two aspects of assessment. While we clearly want to value authentic work and acknowledge student growth, as a school district, we also have serious obligations to conduct student and program assessment responsibly. As we move toward more open-ended and authentic forms of assessment, no one should be allowed to fall through the cracks.

Moreover, as our approaches to assessment move in this direction, serious questions arise about equity and costs. Fortunately, Pacesetter allows our teachers to work with an extensive team of researchers and assessment experts from Educational Testing Service. They are proposing new ways of combining our need to assess students' knowledge with our interest in recording their progress toward valued outcomes.

Unanswered Questions

Many questions about Pacesetter are still unanswered. Present the program to teachers and administrators, and many hands fly up. People want to know:

1. When fewer than half of our students sign up for fourth year math or science, how can we get *all* students to a level where they can take Pacesetter courses in 10th or 12th grades?

2. Particularly in hard financial times, how will we give teachers the time they need to teach and sustain the extra demands of Pacesetter courses?

3. How can we use Pacesetter courses—which are still taught within traditional subject-matter boundaries—to move toward a more integrated high school experience?

4. Pacesetter courses are supposed to be designed for all students. How will we include students with weak academic histories, special education needs, or languages other than English in such demanding courses?

5. The College Board produces other forms of student testing, such as the SATs and the Achievement Tests. How will Pacesetter's more open-ended approach to student assessment affect these other tests?

There are no simple answers. Pacesetter is a "work in progress," just as the College Board is involved in rethinking its mission as a major educational institution. At the turn of the century, it was a tremendous move toward equity to insist that all students be eligible for college on the basis of a common exam. No longer could your last name, and your father's occupation and education, be the gatekeepers to education after high school. A hundred years later, we have learned that equity demands additional tools. We cannot claim to have "done our job" when we have not offered instructional and assessment opportunities that prepare students for college or the world of work. In that light, we are going to have to reinvent our means. Pacesetter provides one laboratory in which to do so. ■

Thomas W. Payzant is Superintendent of San Diego City Schools, 4100 Normal St., San Diego, CA 92103. **Dennie Palmer Wolf** is Director of the Performance Assessment Collaboratives for Education (PACE) Project, and Senior Research Associate, Harvard University, 8 Story St., Cambridge, MA 02138.

"Ensuring Reliable Scoring"

A Chapter in *A Practical Guide to Performance Assessment*, by Joan L. Herman, Pamela R. Aschbacher and Lynn Winters, Association for Supervision and Curriculum Development, 1992

In *A Practical Guide to Performance Assessment,* the authors offer guidance on the creation and use of alternative measures of student achievement. They present a process model that links assessment with curriculum and instruction based on contemporary theories of learning and cognition.

The chapter reproduced here, "Ensuring Reliable Scoring," emphasizes the fact that a fundamental feature of performance-based assessment is its reliance on human judgment. As any trial lawyer will attest, two people viewing the same occurrence or reading the same document often come up with conflicting perceptions or interpretations. Likewise, persons viewing the same behavior on different occasions may arrive at different judgments about that behavior. This chapter is intended to help developers minimize such differences by developing sound scoring procedures.

This work was supported by the Office of Educational Research and Improvement, U.S. Department of Education, Cooperative Agreement Number R117G10027 and CFDA catalog no. 84.117G. Copies of the book may be ordered from:

Association for Supervision and
Curriculum Development
1250 N. Pitt Street
Alexandria, VA 22314
(703) 549-9110

Price: $10.95
ASCD Stock Number: 611-92140
ISBN: 0-87120-197-6

Copyright, 1992 by the Regents of the University of California.
Reprinted with permission of the Regents of the University of California.

Joan L. Herman
Pamela R. Aschbacher
Lynn Winters

BEST COPY AVAILABLE

6

Ensuring Reliable Scoring

A fundamental feature of performance-based assessment is its reliance on human judgment. As any trial lawyer will attest, two people viewing the same occurrence or reading the same document often come up with conflicting perceptions or interpretations. Likewise, persons viewing the same behavior on different occasions may arrive at different judgments about that behavior. The user or developer of alternative assessments must seek to minimize such differences; otherwise the measures cannot be fair, consistent, or valid. Sound scoring procedures help the process.

Understanding the Importance of Reliability and Consistency

The most obvious reason for consistent scoring is equity. To be meaningful, judgments of student performance cannot be capricious. You need to have confidence that the grade or judgment was a result of the actual performance, not some superficial aspect of the product or scoring situation. Was Yuki's grade unduly influenced by her spelling? Did Mark get a better (or worse) grade because his project was graded near the end when you were tired? How was Jamal's grade affected by the fact that another teacher did part of the scoring? What about Corinne? Did she fail the competency writing test this year because the raters were more stringent than last year?

Inconsistency is especially troublesome when the results influence important decisions about students or programs. What grade does Denisha deserve? Should Marta be allowed to take the Advanced Placement English class despite low standardized test scores? Should the school's new math program continue? Even when the results of a single assessment do not carry high stakes, inconsistency means inaccurate scoring. More to the point: inconsistent scoring means the scores have little meaning. If an "A" doesn't consistently represent excellent performance, then what does it mean? The best in the class? The best of a poor lot? Improved effort? If a performance or project receives different scores from different judges, what does each really mean? Which one is accurate? If you apply criteria differently depending on how long you've been scoring, what does the final set of scores mean? What does an individual's score mean?

Achieving Consistency

Equitable and meaningful scoring requires informed and consistent judgment. How do you avoid capricious subjectivity? As we discussed in Chapter 5, having well-defined and defensible criteria for judging student performance goes a long way toward achieving consistent scoring, but there are other conditions that must be met to ensure consistent scoring. First, those making judgments—you, teacher colleagues, the state department of education—must thoroughly understand the criteria in a similar fashion. A consensus among raters about the meaning of the criteria and how they are to be applied builds the foundation for scoring consistency. Second, you need a system for monitoring the consistency of ratings over the period in which performance is being judged. This consistency has several facets. Two or more judges rating the same performance should have general agreement. One judge should rate a particular performance in much the same way regardless of when it is observed—whether during the beginning of the day, somewhere in the middle, or near the end. Judges should rate the same performances similarly on separate occasions. And, the same performances rated on two separate occasions by two different group of judges should be rated similarly. If your scores are used to make high-stakes decisions such as promotion, graduation, or special class placement, you should formally document evidence of scoring consistency.

Professional Development Benefits

The process by which judges learn to apply scoring criteria in a consistent manner can provide a valuable opportunity for professional development. Rater training helps teachers come to a consensual definition of key aspects of student performance. This can lead to a reprioritization of classroom goals as well as insight about the strengths and weaknesses of their students' performances. The scoring process can provide a model for classroom assessment and encourage more collaboration among teachers in the appraisal of student outcomes.

To reap the benefits of consistency and professional growth, you will need good training procedures and a carefully structured rating process. This chapter outlines major considerations in devising and implementing a valid scoring procedure. Although the process we describe has its origin in formal, high-stakes assessments at the district and state level, keep in mind that consistent scoring applies to all forms of assessment, be they classroom grades or college admissions. Decisions about a student can't be valid unless based on reliable information.

Rater Training: A Prerequisite for Consistent Scoring

There are a number of ways to achieve consistency. Our approach emphasizes training raters to a common standard because this approach is efficient and provides teachers with instructionally useful information. Other approaches devote less attention to rater training and consensus-building and rely on multiple judgments of student work to achieve a similar result. As you might expect, the approach you choose depends on your assessment purpose and available resources.

During rater training, judges learn what the scoring criteria mean, what aspects of performance each is intended to capture, and what each of the scale points represents. It is during the training session that you make sure raters apply the criteria consistently to a range of student work samples. This is also the time when raters learn how to record their scores.

Training Manuals

Formal scoring manuals can be very helpful both during and after training. For large-scale assessments, such as yearly district or state testing programs, a scoring manual provides an "institutional memory" of assessment procedures and serves as a useful reference for interpreting scores. For high-stakes classroom assessments, such as Advanced Placement "screening" examinations, or an algebra readiness test, scoring manuals can be useful in discussions with parents or students who want to know how scores are achieved or improved. Typical scoring guides include:

- Fully explicated scoring criteria;
- Examples or models illustrating each score point;
- An abbreviated, one-page, version of the criteria or reference during actual rating; and
- A sample form for recording scores.

You might want to review training manuals from several sources before designing your own rater training. If you are interested in a detailed description of the rater training process, a complete scoring manual developed by the Riverside Publishing Company appears in *Educational Performance Assessment*, edited by Fred Finch (1991). State departments of education are also sources of published scoring manuals.

Training Procedures

Actual rater training is designed to create a consensual understanding of the scoring criteria, provide extensive practice in actual scoring, and, in the case of high-takes assessment, document acceptable levels of scoring consistency (reliability). During rater training, practice scoring sessions provide raters immediate, substantive feedback about their judgments and ample opportunities to ask questions. Raters also come to understand that their job is to make a judgment based on the scoring rubric, not to revise or criticize the rubric and then follow their own inclinations. Without such an understanding, an entire assessment enterprise can be sabotaged.

A typical training session includes:

- **Orientation to the assessment task.** Raters receive an overview of the assessment context, what the results will be used for, who will use them, what directions and prompts the students received, and how the scoring guide operationalizes desired outcomes or processes. It is common to ask raters to actually take the test as a means of orienting them to the scoring task.

- **Clarification of the scoring criteria.** In this phase of training, raters engage in extensive discussion. Both the criteria dimensions and scale values are defined and a range of models provided to exemplify each. Discussion often moves from simpler judgments, such as which samples illustrate high, medium, or low performances, to more difficult distinctions required for assigning numerical scores.
- **Practice scoring.** This is the heart of the rater training process. At first, sample assessments are scored one at a time with discussion following each paper. As raters become more fluent with the scoring guide, they get opportunities to exercise more difficult judgments with problematic (atypical) or borderline assessments.
- **Protocol revision.** During the discussion and practice scoring, raters naturally devise certain rules for dealing with the unanticipated aspects of judgment posed by a particular set of papers and not covered by the scoring guide. For example, when almost every student has misinterpreted the test prompt in the same fashion, rather than to score all answers as "off topic" or "unacceptable," raters may decide to assign scores based on the student-defined task. Or, if many traits are to be scored, raters may decide that different raters should specialize in scoring a few of the traits rather than having all raters score every sample on every dimension.
- **Score recording.** For all assessments, student scores must be recorded in some fashion, on the roll sheet or on summary sheets for a classroom, grade level, or school. Rater training covers the format for recording scores and any special procedures for calculating student scores such as averaging and totalling across dimensions.
- **Documenting rater reliability.** Rater training ends when there is agreement that scorers have reached an acceptable level of consistency, usually rating sample pieces within one point of each other. In order to determine when raters are ready for the real thing, reliability checks are conducted during training. Figure 6.1 provides an example of how to check rater consistency using the percent agreement method.
- **Scheduling Considerations.** How much time will it take to train raters to an acceptable level of agreement before letting them judge student work? It depends on:

  — How experienced your raters are.
  — Whether they are familiar with your scoring criteria.

- —How quickly raters come to consensus about the meaning of the criteria.
- —The complexity of the scoring criteria, and the quality of the work to be judged—with borderline work being the most difficult to assess quickly.

We have found that it takes about three to four hours to train raters to use a holistic or simple (two- to four-trait) analytic scale. More complex scales can require up to a full day of training.

Rater fatigue is an important factor in scoring; we consider a six-hour session a full day's work. You should also schedule time for retraining or refreshing raters at the beginning of each new scoring day, and certainly for any changes in topics or tasks that use the same scoring

Figure 6.1
Calculating Rater Agreement
(Three raters for two papers)

	Is Rater in Perfect Agreement with the Criterion Score?			**Is Rater in Agreement with the Criterion Score, Plus or Minus 1 Point?**		
Rater	Paper #1	Paper #2	Rater's Average Agreement	Paper #1	Paper #2	Rater's Average Agreement
Linda	yes	no	50%	yes	no	50%
Robert	no	no	0%	yes	yes	100%
Ella	yes	yes	100%	yes	yes	100%
Total	67% = yes	33% = yes	50%	100% = yes	67% = yes	83%

Figure 6.1 illustrates the case in which three raters are asked to rate two criterion papers after some training. According to the results in the figure, Linda agrees with the criterion score for paper 1 but not for paper 2; in fact, for paper 2 she is not even within one point of the criterion score. Robert is not in perfect agreement with the criterion scores on either paper 1 or paper 2 but is in agreement plus-or-minus one score point on both papers. Ella is in agreement all the time and is ready to rate student work. Robert and Linda probably need a little more training. Paper 2 causes more problems for raters than paper 1, so further training should focus on distinguishing the criterion score from neighboring scale points. In reporting these results you could say, "On average, raters obtained perfect agreement with criterion scores 50 percent of the time, and reached ±1 agreement 83 percent of the time.

criteria. In high-stakes assessment, retraining often takes place after any lengthy breaks such as lunch.

Training Paper Issues

Because rater training provides a dry run for actual scoring, it behooves you to anticipate as many possible sources of rater disagreement as possible before rater training and to build opportunities into the training papers for eliciting disagreement and discussing it. For example, the syntactical constructions used by non-native English speakers raise issues related to balancing content with communication concerns. You should also deal with handwriting and legibility issues or aesthetic quality concerns in visual and performing arts. Finally, you want to be sure that the sample papers you select for training represent not only each point on the score distribution but also the entire range of student performance likely to be encountered in scoring. The natural human tendency is to grade normatively. The better work samples from a set of relatively poor papers may receive higher scores than they would were they part of a stack of relatively good papers. The reverse can also be the case. This tendency should be discussed during rater training with examples provided so that the scoring criteria maintain the same meaning across different sets of papers and different scoring occasions.

Obtaining Sample and Check Papers

Because a wide array of sample work is needed to guide raters, you should collect samples from a diverse group of students. Pick work from a field-test, a previous assessment, or from the actual assessment. To identify appropriate training and check papers, a group of "experts"—teachers from the grades and subjects involved who are familiar with your scoring criteria—can be quite helpful. They can select examples that illustrate the range of responses, from clear to borderline, for each score point so that raters will be trained to handle all situations. If several prompts or tasks are used in the assessment, examples need to be drawn for each. If you are using age-related scales across grade levels, you need examples to illustrate each age level. It is also useful to prepare comment sheets explaining how the specific aspects of each piece of work represent criteria for a particular score. The expert group can then identify samples that will be used for (1) training discussions,(2) practice, and (3) checking consistency.

Score Recording Concerns

You need to provide raters a method for recording student scores. In your own classroom, you might simply record scores at the top of the student's paper and then in your roll book. Some teachers use the scoring criteria as a feedback sheet for students. They circle deficient areas or note strengths using the descriptors on the guide. The same process can be used to create a classroom profile on one master scoring guide.

In more formal assessment settings, score sheets become a matter of public record and are used to provide feedback to teachers and others. Data analysts also use them to calculate test statistics. In these instances, raters are often given machine readable documents for "bubbling" in student scores as well as other important information such as the school, district, student, and rater identification numbers and the code numbers for topic or task and date. Whenever you have two or more raters scoring student work, you'll need to remind them not to indicate scores, comments, or corrections on the sample itself. You don't want a subsequent rating influenced by their comments.

Reliability Issues

The purpose of rater training is to create consistent, reliable scoring procedures. Thus, a method of determining if raters are consistent should be built into the training period. Many strategies for checking rater reliability exist. One commonly employed approach is to prepare in advance and score a set of ten or so "reliability check" papers representing the range of student performance. Ask the raters to score this same set and compare their judgments with you or others who are trusted assessors. Reasonable agreement with both the expert judgments and with each other suggests that raters are ready to score actual student work.

What constitutes reasonable agreement? You can ask that all raters be in exact agreement before you consider them reliable, or you can use the less stringent "plus or minus one" rule, which is fairly common and says that raters are "in agreement" if they agree within one scale point, "plus or minus." For example, if the score on a particular reliability-check sample is a "3," anyone who gave it a rating of "2", "3," or "4" is considered to be on target.

Regardless of the target level of agreement you choose, when you train raters, the goal is to have them apply the scoring criteria exactly as intended, not to within one scale point of the target score. When a rater has difficulty applying the criteria exactly as intended, you should spend time during training discussing the practice papers, criteria, and decision rules for applying the criteria in order to bring the rater up to an acceptable level of consistency. However, some raters may not be able to adjust their internal criteria to match the scoring guides. These aberrant scorers should be dismissed or assigned to other tasks during actual scoring.

In addition to deciding how close ratings should be to establish consistency, you need to think about how often they need to be in such agreement. If you are asking for exact agreement, which can be difficult to obtain, your criterion for reliability may be less stringent than if you are using the "plus or minus one" rule. At CRESST, we often ask that raters agree with the experts at least 90 percent of the time on each scoring dimension when using the "one point off" guideline. The guideline for exact agreement could drop to 75 to 80 percent under the more stringent condition. The actual percentage of agreement varies depending on the assessment purpose and stakes involved.

Regardless of how you define "rater agreement," the purpose of reliability checks is to ensure that student scores aren't the result of capricious judgment, one of the most commonly cited arguments against performance assessment. Consider the classic study conducted by Paul Deidrich (1963) at the Educational Testing Service in which the same essay was assigned an entire range of scores by a group of raters. What most don't remember about this study is that acceptable levels of rater agreement were obtained when the judges (1) were drawn from the same discipline, (2) used explicit scoring criteria, and (3) participated in a training session.

Ensuring Equitable Judgments During an Actual Scoring Session

Maintaining Consistency

Documenting rater consistency during training is simply the first step toward creating a fair, equitable scoring process. Because the purpose of rater training is to develop rater consistency, you need to monitor rater scoring patterns during the actual scoring process as well. Research shows that raters have a tendency to drift away from formal criteria to their own, more idiosyncratic views (Quellmalz and Burry 1983). Human judgments and expectations are shaped not only by formal standards, such as scoring criteria, but also by their prior experience and the actual range of performance currently being assessed. If the entire set of performances appear to be relatively "poor" according to the objective criteria, raters develop a tendency to shift the criteria downward so they can award higher scores to the "best of the worst" papers. As a teacher, you too have perhaps been aware that your standards and expectations for students change during the grading process. You modify your ideas somewhat after looking at several pieces of student work. For this reason, training sessions need to include a large sample of papers and the entire range that might be encountered during actual scoring.

For classroom assessment purposes, you can check your consistency by stopping midway and rescoring some of the first student work you scored. When you are scoring several different dimensions or topics, you can score all work on one dimension or related to one topic at the same time, then go back and score for other factors. Scoring all papers several times, once for each different dimension or topic, is often quicker than going through individual papers for everything at once and applying multiple criteria or reading different kinds of responses. Your scoring pace also increases as you become familiar with the criteria.

For school-level, larger-scale, or high-stakes assessment, you'll want to build in more formal rater consistency checks. For essay scoring this is sometimes done by burying pre-scored common check papers at designated intervals in each rater's stack of papers. The scoring director then checks raters on the common paper and works with those who have drifted away from a consistent application of the scoring guide. Another method is to conduct mini-training sessions first thing in the morning or right after lunch. Raters score a common set of check papers, much as they did in training. Those who have drifted from the preset standard (exact agreement; plus or minus one point) participate in a review session and are rechecked before being allowed to continue scoring.

An additional consistency consideration in large-scale assessment relates to lack of bias in rater judgments. You need to be sure that raters working together don't form subgroups who agree with each other but not all the other participating raters. To avoid this, break up rater groups at periodic intervals and have second ratings of papers/work done by raters assigned to other tables or physical locations.

Managing Logistics

Although achieving consistent judgment is the overriding concern of scoring, conducting a scoring session involves a number of logistical and technical issues. Scheduling is one of the most fundamental concerns in planning a scoring session. As people tend to tire in the afternoon and rate more slowly, you might consider scheduling your rating sessions early and avoiding the late afternoon. Access to a copy machine enables you to address any unanticipated shortages of rating materials or to reproduce papers that require discussion during the rating session. Further, rating is an intense activity; provide frequent breaks and snacks (lots of fruit and carbohydrates, little sugar). The scoring area itself should be quiet and comfortable with ample room for raters to accommodate the work to be reviewed. A rater's nightmare is to work in the gym on folding chairs and tables at 3:30 on a hot May afternoon during band practice.

Another concern is managing the flow of papers or other student products. In large-scale assessments, each table of scorers should have their own leader whose sole duty is to manage the paper flow and monitor rater consistency. Our experience suggests that bundles of student work that take about one hour to rate are easier for raters to handle than individual pieces. The number of pieces in each bundle will vary with the nature of the task and the complexity of the scoring scheme. In writing assessments, for example, sets often consist of fifteen to twenty-five papers, whereas a bundle of portfolios might include only four to six. Regardless of how work is bundled, individual pieces must be randomly assigned to bundles and bundles randomly assigned to raters so that no systematic scoring effects occur. For formal assessments, both raters and students should be assigned identification numbers to guard against bias and protect privacy.

You'll need to decide whether to mix different grade levels or different topics together in the same scoring session. Generally, this is not done unless the purpose of an assessment is to compare students at different grade levels on the same scoring scale. In large-scale assessments, different topics are either assigned to different rater groups or scored separately from each other with a session of refresher training preceding the topic change.

Another concern that can cause problems later if not monitored carefully is ensuring that scorers are recording required information properly. Were all identification numbers bubbled in along with the scores? Were scores recorded for all papers rated? Do all students have scores? The list is extensive. Try to anticipate what can go wrong and devise strategies for either preventing it from happening or for fixing it.

Ensuring Technical Quality

Advice on all the technical decisions you have make to ensure scoring accuracy and equity is beyond the scope of this book and in fact constitutes a psychometrician's career. If you are assessing for a high-stakes decision, especially if that decision can get you sued, disparaged on page one of your local newspaper, or called before the board of education, you may want to bring in a technical consultant to structure your scoring process and help you document the reliability of student scores. Following are some of the questions you need to address:

How many raters are needed? This, of course, depends on how much work is rated, how many ratings each piece will receive, how long it takes to rate each piece, and how many days are available for scoring. Holistic scoring of one-to-two page essays generally goes quickly, sometimes as quickly as a minute a paper. A complex analytic rating on longer pieces can take four to five minutes per paper. Portfolios can take longer still. As for the number of days, our experience suggests raters can get quite burned out after four or five days.

How many scores per paper? Effective training and vigilant monitoring of the scoring process can eliminate much of the need to do multiple scoring of the same dimension of student work. Multiple raters are needed for each paper when raters are inexperienced or there is little evidence that raters are using the same criteria and standards in making their judgments. The need for multiple scores depends on your assessment purpose. The more serious the consequences, the more important it is that you document consistency. Our experience suggests that no more than two raters are needed for any piece; the ratings can be summed or averaged to provide a final score. A third opinion can be called in for difficult cases, such as the occasional nightmare paper that draws both the lowest and highest score.

In some situations, one score is sufficient for a majority of the pieces. Consider a situation in which selection, placement, or other critical decisions about individual students will be made based on some prespecified standard or cut score. If your training and scoring check papers show that raters are consistent, the only papers requiring two or more ratings will be those borderline papers falling around the passing score. Because rating is an expensive process, you will need to balance reliability concerns against those for cost and efficiency.

How are papers scored for evaluation purposes? If student scores will be used for program evaluation rather than individual assessment, a reliable estimate of an individual student score is less critical than the average score for the task. Most pieces of work can be read only once, and your reliability evidence can be obtained on a sample of work (perhaps 20 percent), which is rated by two or more raters. If you are using student samples to evaluate a program and don't have to provide individual scores to teachers, it is more efficient to score a randomly selected sample of student work. Your technical consultant can advise you about sample size and the appropriate manner of selection.

Providing Evidence of Reliability

For high-stakes assessments, you need to formally document the consistency and reliability of your scoring process. Plan to invest in the services of a technical expert in advance of the scoring to ensure that you have an adequate scoring design, that you are collecting suitable evidence, and that your data are appropriately formatted to ease data analysis.

The following are some relevant sources of evidence:

- **Results of the qualifying check after training.** Plan to report on what agreement level was required. What proportion of your raters passed on the first try? What was the average level of agreement among those passing?
- **Results of the consistency check during scoring.** Plan to report on what agreement level was required. How many and when were the checks made? What proportion of your raters passed without remediation? What was the average level of agreement on the checks?
- **Inter-rater reliability results for student work scored by more than one rater.** Percentage agreement among raters and generalizability coefficients are two frequently used techniques. Each of these is calculated separately for each scale you use. As a guide, you need to double score at least 20 percent of your student samples to get sufficient evidence, and if more than two raters are involved, you need to consult a statistician for help with a balanced design specifying which raters are to score which pieces of student work.

  What level of agreement or reliability is high enough? Of course the answer is: it depends on the decisions you are making. The more critical or restrictive the consequences are, the more reliable your scores need to be. In general, reliability coefficients of .70 and above are considered respectable. Coefficients of .90 and above are not uncommon with standardized multiple-choice tests, and large-scale direct writing assessments.

- **Rater consistency across years.** When you want to be sure that your rating scale is consistent from year to year—for example, when results are being used in state assessments to track trends over time—you need to include with this year's scoring a sufficient sample of student work from last year's scoring. Agreement in scores assigned can then be checked, and if necessary, statistical adjustments can be made for differences.
- **Rater consistency across different locations or different groups of raters.** Similar to checking consistency across years, if student work is to be scored at a number of different locations or by different groups of raters, you need to check on the consistency of these different groups. For example, a state might convene four regional workshops to score its hands-on science assessments, or a district assessment might require each school to score its own students' work. One way to check for consistency would be to seed the work scored by each group with a common set of work. At scoring site one, for instance, scorers would assess student work assigned specifically to site one plus the common set; site two scores would assess student work assigned specifically to site two plus the common set and so forth. Scores on the common set can then be checked for consistency.
- **Inter-rater consistency.** This is the degree to which one rater remains consistent over time. Check for this by having raters score the same piece more than once at different points in the scoring process.

Checking the Reliability of Your Rating Process

As a summary of many of the issues covered in this chapter, use the following checklist to see if your scoring procedures are sound and reliable. Do you have:

[ ] documented, field-tested scoring guide
[ ] clear, concrete criteria
[ ] annotated examples of all score points

[ ] ample practice and feedback for raters
[ ] multiple raters with demonstrated agreement prior to scoring
[ ] periodic reliability checks throughout
[ ] retraining when necessary
[ ] arrangements for collection of suitable reliability data

References

Baker, E.L., P.R. Aschbacher, D. Niemi, and E. Sato. (1992). *CRESST Performance Assessment Models: Assessing Content Area Explanations.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Deidrich, P.B. (1963). "The Measurement of Skill in Writing." *School Review* 54: 584-592.

Finch, F. (1991). *Educational Performance Assessment.* Chicago: Riverside Publishing Company.

Quellmalz, E., and J. Burry. (1983). "Analytic Scales for Assessing Students' Expository and Narrative Writing Skills." (CSE Resource Paper No. 5). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Students Testing.

CRESST Performance Assessment Models: Assessing Content Area Explanations

This handbook presents a performance-based approach to assessing students' understanding of subject matter content. It is based on years of research conducted by the Center for Research on Evaluation, Standards, and Student Testing (CRESST), funded by the U.S. Department of Education's Office of Educational Research and Improvement. The handbook includes: (1) a concise model of alternative assessment for those who need to develop similar assessments on their own; (2) examples of successful CRESST assessment materials; (3) an effective scoring rubric for performance assessments applicable to a variety of topics; and (4) useful benchmark papers.

Four parts of the handbook are reproduced here:

- The Table of Contents
- The Introduction
- Chapter 1: Overview of CRESST Research
- Sample student assessments in chemistry

The handbook was written by Eva L. Baker, Pamela R. Aschbacher, David Niemi and Edynn Sato with support from the Office of Educational Research and Improvement, U.S. Department of Education, Cooperative Agreement Number R117G10027 and CFDA catalog no. 84.117G.

Copies are available for $10 from:

UCLA
CRESST
Graduate School of Education
405 Hilgard Ave.
Los Angeles, CA 90024-1522

Reprinted with permission of the Regents of the University of California.

CRESST PERFORMANCE ASSESSMENT MODELS:

Assessing Content Area Explanations

Eva L. Baker, Pamela R. Aschbacher,
David Niemi and Edynn Sato

April 1992

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

Table of Contents

Introduction 1

Chapter 1

Overview of CRESST Research 3

Chapter 2

Guidelines for Using CRESST's Model for Assessing Explanation 7

Chapter 3

Sample Assessment Materials for Students 14

Chapter 4

Specifications for Developing Assessment Materials 30

Chapter 5

Rater Training, Scoring and Reporting 39

Chapter 6

Sample Training Materials 64

Introduction

Purposes

This handbook presents a performance-based approach to assessing students' understanding of subject matter content. It is based on years of research conducted by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), funded by the U.S. Department of Education's Office of Educational Research and Improvement (OERI). The purposes of this handbook are to:

- provide one model of alternative assessment for those who need to develop similar assessments of their own;
- introduce successful examples of CRESST assessment materials; and
- facilitate research on other alternative assessments.

The materials which follow are the result of our five year research effort designed to explore the development of alternative assessments in history. To summarize, the project has attempted to find ways to score the content quality of essays in history. Using the writing of expert historians as the basis of scoring criteria, we have developed techniques for measuring the deep understanding of history and for scoring student work reliably. Our work has been conducted using students from grades 8 through 12 and has been expanded to other content areas as well (economics and science).

Large-Scale Assessment

These assessment tasks are consistent with cognitive learning theory. They include recalling prior knowledge in a content area, reading primary source documents containing new information, and writing an explanation of important issues that integrates new and prior information.

Our assessment judges student understanding on the basis of six scales, including the use of concepts and facts, the avoidance of major misconceptions, and the quality of the argument presented. The scales were developed from studies of expert and novice performance. We have used this assessment approach to research a number of technical issues in performance assessment and have demonstrated the reliability, validity, and generalizability of this technique.

Reliability, validity, and generalizability

We believe that this assessment could be useful for both large-scale accountability and diagnostic improvement of instruction. Typically, measurement experts have argued that accountability and diagnosis should be conducted with separate kinds of assessments. But for practical, economic, and conceptual reasons, we argue that they can be merged into a single measure, with different methods of reporting the data for different purposes.

Inside you will find background information on our CRESST performance-based assessment, examples of assessments for secondary level history and chemistry, and specifications for duplicating our technique with other topics and subject matter areas. We also describe our rater training process, scoring techniques, and methods for reporting results.

Interested users may contact CRESST at (310) 206-1532 for copies of additional materials, assistance using them in an assessment program, help in developing assessments for new topics, or for technical information about the rating scales.

Chapter 1

Overview of CRESST Research

Cognitively Sensitive Assessment

Tests should measure significant learning in a way that supports desired performance. This simple concept should lead us, as educators, to a reversal of our present use of standardized tests which fail to measure deep understanding of student learning. Instead of having tests constrain instruction, assessment procedures should build directly on learning.

Assessment procedures should build directly on learning

Despite the widespread interest in alternative assessments, there has been relatively little research on the design and technical quality of such measures. CRESST began conducting research on history performance measures in 1988. Focusing on both explanation and knowledge representation skills, we have attempted to develop a better method for validly assessing secondary students' deep understanding of content areas such as history.

Many current performance assessments are developed with minimal design constraints because clearly acknowledged technology does not exist for performance task design. Developers seem to focus on a few limits when they create new assessments. One set of constraints concerns logistical issues, such as assessment time and availability of materials. Another emphasis has been on the surface characteristic of the task, that it exhibits motivational or "authentic" attributes of the assessment.

Teachers and other developers want assessments that capture the imagination of students, intrinsically motivate, and if possible, demonstrate relevancy to real-world demands and expectations. Far less attention has been paid to design constraints focused on increasing the technical quality and the economic feasibility of the resulting assessments.

Assessments that capture the imagination of students

CRESST's research assumes that a desired goal of performance assessment is the generation of "comparable" tasks for estimating student achievement. Our approach has sought to produce comparability by designing it at the outset rather than adjusting findings *post hoc* through scaling and statistical equating. Specifications to control cognitive demands of the task, the structure of the assessments, and the generation and application of scoring rubrics have been thought to produce performance that showed less variability from topic to topic than tasks created with fewer design constraints. In our attempt, we have tried to control both rater and score reliability.

Control both rater and score reliability

Our history performance tasks, which have evolved over time, require students to engage in a sequence of assessed steps—taking a minimum of one-and-a-half hours per topic. First, students are assessed on their relevant background knowledge of the particular historical period. This measure consists of a 20-item, short-answer test with questions to measure student knowledge of historical principles and specific events pertinent.

Next students are provided with opposing viewpoints in primary source text materials, typically letters or speeches of historical figures. Finally, students are asked, in a highly contextualized set of directions, to write an essay that explains the positions of the authors of the texts, and to draw upon their own background knowledge for explanation. In some studies we have given students optional resources to read, or have asked students to prepare HyperCard or concept map representations of the key knowledge, principles and relationships in the text materials (Baker, Niemi, Novak, & Herl, in press).

An essay that explains the positions of the authors

CRESST conducted a series of studies to determine how scoring rubrics should be developed, and the best strategy relied on looking at differences between expert and novice performance (Baker, Freeman, & Clayton, 1991). The essay scoring rubric consists of six dimensions, a General Impression of Content Quality scale (focused on the overall quality of the content understanding), and five analytic scales:

Five analytic scales

- Prior Knowledge (the facts, information, and events outside the provided texts used to elaborate positions);
- Number of Principles or Concepts (the number and depth of description of principles);
- Argumentation (the quality of the argument, its logic and integration of elements);
- Text (the use of information from the text for elaboration);
- Misconceptions (the number and scope of misunderstandings in interpretation of the text and historical period).

Each of the above dimensions is scored on a 0-5 point scale.

Sets of history assessments have been developed

History experts and high school teachers have been involved throughout the study as co-designers, reviewers, and raters of the assessment. So far, six complete sets of history assessments have been developed: two on the Revolutionary period; one on the Civil War; two on 20th century immigration; and one on the Depression Period. These tasks connect to the *California History-Social Science Framework* (1988). Replications in the areas of science (Baker, Niemi, Novak, & Herl, in press) and economics (Baker, 1991) have been conducted to assess the utility of the scoring rubric for explanation tasks in other content areas.

What CRESST Has Learned

CRESST research

Over the past several years of research on this project, CRESST has:

1. developed a valid scoring scheme for assessing deep understanding of history, generalizable across topics;
2. developed rater training procedures that produce reliable and valid scoring of student tasks in a limited period. The scoring rubric makes strong cognitive demands of the raters;

What CRESST has learned

Fairness, generalizability, cognitive complexity, content quality, reliability, cost, and efficiency

3. built a task structure that reduces score variability so that fewer topics can be used to derive reliable scores for individual students. This technique is more efficient than found in most comparable studies. These relationships are all the more startling because of the lack of preparation and motivation among our students;

4. distinguished between assessment purposes and the utility of overall score and subscores;

5. found gender differences in this small sample, favoring females;

6. found supportive data for the validity of our measures in grade point average (GPA) and a scale measuring student effort;

7. systematically addressed validity criteria (Linn, Baker, & Dunbar, 1991) in our research studies: the criteria addressed include fairness, generalizability, cognitive complexity, content quality, reliability, cost and efficiency. We are in the process of conducting studies of transfer and designing research to assess the meaningfulness of tasks to students.

For additional details on the background, development and methodology of this research, please contact the CRESST office.

References

Baker, E.L. (1991). *Foundation for Teaching Economics evaluation report*. Los Angeles, CA: Author.

Baker, E.L., Freeman. M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition*. Englewood Cliffs, NJ: Prentice-Hall.

Baker, E.L., Niemi, D., Novak, J., & Herl, H. (in press). Hypertext as a strategy for teaching and assessing knowledge representation. In S. Dijkstra (Ed.), *Instructional models in computer-based learning environments*. Berlin: Springer-Verlag.

California History-Social Sciences Framework. (1988). Sacramento: California State Department of Education.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21.

Example 3.4
Prior Knowledge Measure
Chemistry

Name__________________

How Much Do You Know About Chemistry?

Directions: This is a list of terms related to high school chemistry. In the space after each term, write down what comes to mind drawing upon your knowledge of chemistry. A brief definition would be acceptable, or a brief explanation of why that law, principle, concept, or procedure is important in explaining chemical phenomena. If a term is general, give both a general definition as it relates to chemistry and a specific example to show your understanding, if you can.

Good Example: PERIODIC TABLE — An arrangement of chemical elements based on the order of their atomic numbers. Shows variation in most of their properties. Shows a natural division of elements into metals and nonmetals, inert gases, atomic weights.

Do not define the term by simply restating the same words.

Bad Example: ELECTRON LEVEL — The level of the electron.

Even if you are not sure about your answer, but think you know something, feel free to guess.

There are probably more items here than you will be able to answer in the time given. Start with the ones you know best, and work quickly so that you can answer as many as possible. Then go back and answer the ones of which you are less sure. Do not spend too much time on one specific item.

1. density ____
____
____

2. solubility test

3. conductivity

4. chemical reaction

5. base

6. nucleus

7. deductive reasoning

8. conservation of energy

9. precipitation

10. fructose

11. hypothesize

12. empirical formula

13. acid ____________________

14. experimental control ____________________

15. gas laws ____________________

16. compound ____________________

17. ion ____________________

18. indicator ____________________

19. quantitative analysis ____________________

20. hydration ____________________

Example 3.5
Chemistry Demonstration: Soda Task*

As an introduction to chemical analysis, a high school chemistry teacher performed an experiment for her class. This is a description of what she did.

"I have two samples of soda," she told the class. "One is regular soda containing sugar and the other is diet soda which contains an artificial sweetener. I'm going to identify each sample as diet or regular by doing some chemical tests. As in any chemical testing, I won't allow myself to taste the samples but will base my decision solely on the chemical and physical properties of the two samples as determined by the tests."

She began by labeling the samples A and B to help her keep track of the sample she was testing. She then proceeded by saying, "Since we've been studying the properties of many different kinds of substances, we know that we often can identify an unknown substance by performing physical and chemical tests on the substance and observing reactions. For example, acids turn certain solutions pink, while alkalis turn them green, and neutral ingredients fail to change the color of the solution. Keeping in mind the chemical properties of sugar, I'm going to conduct the following tests: the yeast test, the benedict solution test, a test using sulfuric acid, a solubility test, a test using salt, and a residue test."

Her first test was the yeast test. She poured equal amounts of each soda into separate test tubes and labeled them A and B respectively. One soda reacted with the yeast to give off a distinctive odor as well as gas bubbles, and the other did not react in the same way.

Next she used a benedict solution test. She began by pouring the indicator (benedict solution) into three test tubes. She then added a portion of soda A to one test tube and an equal portion of soda B to another test tube, making sure to note on each test tube which soda was added. The third test tube was a control: nothing was added to the indicator in this test tube. She waited, knowing that some substances take a while to react with the indicator. Comparing the two test tubes containing soda with the control, she pointed out that a reddish precipitate had formed in one of the test tubes.

For her next test, she mixed sulfuric acid with each of the sodas, handling the acid with extreme caution. She began by heating the sodas so that most of the liquid evaporated. Then as she added the sulfuric acid to each sample, she noticed that the acid reacted with one of the sodas to form a gooey brown substance.

To conduct the solubility test, she poured 100 ml of soda A and 100 ml of soda B into separate beakers and gradually added equal amounts of sugar to each soda. She stirred the sodas and waited 15 seconds to see if the sugar dissolved. She found that more sugar dissolved in one soda than the other.

Next she prepared new samples containing equal amounts of each soda and added equal amounts of salt to each sample. She noticed that as salt was added, one soda fizzed more than the other.

Finally, for the residue test, she placed 30 ml of each soda in separate test tubes, placed both tubes over a Bunsen burner and heated them until 15 ml evaporated from each. She noticed that more residue was left in one of the test tubes.

Upon completing the various tests, she made a chart of the results which looked like this:

	A	B
Yeast test	distinct odor gas bubbles	no odor no bubbles
Benedict solution test	reddish precipitate	no precipitate
Sulfuric acid test	produced a gooey brown substance	no gooey brown substance
Solubility test	not much sugar dissolved	a lot of sugar dissolved
Salt test	not much fizzing	a lot of fizzing
Residue test	a lot of residue	not much residue

The teacher ended her demonstration by saying, "With your knowledge of the properties of sugar and the results of the tests, you now can determine which of these sodas is the regular and which is the diet."

**This task was adapted with permission from one developed and tested by the Connecticut State Department of Education.*

The CRESST Line

Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing

Special Portfolio Issue

Includes:

"Portfolios and Accountability," by Eva L. Baker and Robert L. Linn

"Portfolios as Worthwhile Burdens?" (some results from Vermont's portfolio assessment program) by Ron Dietel

"Records of Achievement: Lessons from the United Kingdom"

CRESST Line is produced with funding from the Office of Educational Research and Improvement, U.S. Department of Education, Cooperative Agreement Number R117G10027 and CFDA catalog no. 84.117G.

To be placed on the *CRESST Line* mailing list, write:

Ron Dietel/Mail List
CRESST Line
UCLA Graduate School of Education
405 Hilgard Ave.
Los Angeles, CA 90024-1522

Reprinted with permission of the Regents of the University of California.

Fall 1992

The CRESST Line

NEWSLETTER
of the
National CENTER
for RESEARCH on
EVALUATION, STANDARDS,
& STUDENT TESTING

Special

PORTFOLIO

Issue

Inside CRESST LINE

From the Directors .... *page 1*
CRESST Products .... *page 2*
Portfolios as Worthwhile Burdens ..... *page 3*
New Assessment Book .... *page 5*
Records of Achievement .... page 6
CSE/CRESST Reports .... *pages 9-11*
CRESST Waves .... *page 12*

From the Directors —

PORTFOLIOS AND ACCOUNTABILITY

by Robert L. Linn and Eva L. Baker

"All things to all people" could be a reasonable subtitle for performance assessment. This new type of achievement measurement is promoted as revitalizing teaching, reforming curriculum, and motivating students. Performance assessment is claimed to be useful for evaluating programs, improving instruction, comparing districts, and evaluating university and job applicants. Tomorrow's news will probably report it lowers cholesterol as well.

Eva Baker

Robert Linn

The hyperbolic expectations associated with performance assessment have created a situation paradoxically analogous to previous testing practices now in the throes of mass repudiation. Commercial test producers, you may recall, consistently reminded users that the purposes of their traditional (i.e., multiple-choice) measures were limited. It was the presumably school users who expanded the applications of these achievement tests from reports of individual achievement to measures of accountability. When high stakes became associated with performance, so the story goes, then instruction inappropriately focused on the test. Validity of interpretation suffered, some believe standards were lowered, and in any event, these tests became regarded by many as an educational scourge.

Now it appears that the educational community is on the verge of repeating at least the syntax if not the details of the same mistake. Its vehicle may be portfolio assessment, considered one of the most appealing manifestations of performance-based assessment. Portfolios of student accomplishment allow the collection of a cumulative record of a student's growth. Following the metaphor from the visual arts, a portfolio can include a selection of the student's prized efforts, a display of virtuosity, and even progression through developmental stages, her own Blue Period, for example.

Portfolios conceived as intensely personal portraits of accomplishment would seem to have a number of desirable consequences. Students would have to become active in the process and choose

(continued on page 8)

NEW PORTFOLIO VIDEOTAPE!

"Just collecting children's work does not accomplish the tasks of setting standards and measuring learning."

Excerpt from the new CRESST videotape
"Portfolio Assessment and High Technology"

Portfolios, as effective assessment, should include clearly defined student standards designed in a collaboration process that includes classroom teachers.

This statement is one of the key messages in "Portfolios and High Technology," a recently released videotape produced by UCLA's National Center for Research on Evaluation, Standards, and Student Testing (CRESST), with support from Apple Classrooms of Tomorrow (ACOT)sm. Join the CRESST research staff, including Eva Baker, in this 10-minute videotape as they examine the important issues and uses of portfolio assessment including:

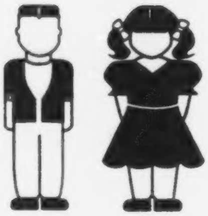
- Theoretical needs for improved assessment;
- Student use of portfolios in the classroom;
- Improved teacher motivation through the portfolio process;
- Teacher workshops that help to define and select standards;
- Involvement of parents in the portfolio process; and
- Use of technology to promote good writing.

Demonstrating effective use of classrom technology, the tape shows students engaged in a variety of activitites including:

- Selecting best pieces of classroom work; and
- Development of computer-based portfolios.

Useful for school districts, principals, and teachers interested in starting or improving their own portfolio programs, this tape will also interest researchers who want more information about the latest CRESST research into portfolio assessment. Although this videotape emphasizes technology, the content is applicable to nearly all portfolio programs.

Cost of the tape is $10.00 and it may be ordered on page 11.



For those interested in portfolios, the following technical report will be of special value.

Writing Portfolios at the Elementary Level:

A Study of Methods for Writing Assessment

Maryl Gearhart, Joan Herman,
Eva Baker, & Andrea Whittaker

CSE Technical Report 337, 1992
($4.00)

This study examines the feasibility of using student portfolios to evaluate writing competence. The authors found that analytic portfolio ratings showed promising levels of measurement quality, but differences in assessed level of performance emerged when portfolio scores were compared to other assessments. Qualitative analyses of the scoring process revealed significant design challenges, particularly in devising portfolios that reflect classroom instruction yet are sufficiently uniform to permit meaningful comparisons within and between classrooms and schools.

Please use the order form on page 11 or call Kim Hurst at CRESST (310) 206-1532, for a complete list of CRESST technical reports.

PORTFOLIOS AS WORTHWHILE BURDENS?

by Ron Dietel

The "bad" news: Portfolios are indeed a major time and resource burden on schools, especially on teachers.

The "good" news: The instructional and motivational results from portfolios may lead to important changes in classroom practices.

Interim results from two research studies provide evidence to support such conclusions, at least for *large-scale* portfolio systems. In a new report, ***The Vermont Portfolio Assessment Program: Interim Report on Implementation and Impact, 1991-92 School Year,*** teachers and principals reported that implementing the portfolio program required considerable time and effort. Even so, many of them felt that positive classroom effects were the result.

Authors of the CRESST/RAND report, Dan Koretz, Brian Stecher, and Edward Deibert, have been evaluating the Vermont portfolio program for almost two years. Vermont was the first state to make portfolios the backbone of a statewide assessment system.

The researchers say that support for the Vermont portfolio program, despite tremendous demands on teacher time, is widespread. "Perhaps the most telling sign of support for the Vermont portfolio program," wrote the authors, "is that [even in the pilot year] the portfolio program had already been extended beyond the grades targeted by the state."

In a second statewide portfolio assessment project, the Michigan Employability Skills Portfolio, teachers and school officials have also reported increased demands on their time. As noted by Michigan educators, one of the key issues for portfolios is their overall cost factor—expenses such as duplication costs, storage space, time for training teachers, and time for scoring portfolios. Despite the cost issue, teachers in Michigan expressed enthusiasm about portfolios.

"When students ask me why do I need to learn this, I have a real answer now..."

"When students ask me, as they had in the past, why do I need to learn this, I have a real answer now," said Rita Kirby, a teacher from Ithaca, Michigan. "I tell them that the kinds of skills that they are developing through the portfolio [program] are skills that are going to be serving them for the rest of their lives."

Background-Vermont

Unlike most other states, Vermont had no statewide educational testing program until a portfolio system was selected in 1988. Since then, Vermont has been developing a "cutting edge" assessment program, the centerpiece of which are portfolios of students' work and "best pieces" drawn from them. The Vermont assessment program currently includes mathematics and writing portfolios in grades 4 and 8 and will eventually encompass a broader range of subject areas.

Vermont's use of the portfolio results will be limited compared to some state and national proposals for accountability. Although schools and principals may use the results for assessing individual student skills, the state will use the results only as a barometer of school and district movement towards state goals of instructional reform. In mathematics, for example, Vermont wants students to increase their problem-solving skills, understanding of patterns and relationships found in mathematics, and communication of mathematical concepts. Early results from the Vermont research indicate that teachers, in response to the Vermont assessment program, are indeed spending significantly more instructional time on these specific areas.

Background-Michigan

In the 1980s, Michigan's economy was suffering from a serious loss of manufacturing jobs. To improve the employability skills of its workforce, the [Michigan] Governor's Commission on Jobs and Economic Development convened the Employability Skills Task Force in 1987. The task force's mission was to identify "skills" essential to new employees entering the workforce.

After a comprehensive survey of over 2000 businesses, the task force published the Employability Skills Profile, expanding on three previously identified areas of needed student achievement: academic, personal management, and teamwork skills. Portfolios, at that time already growing in popularity in some Michigan schools, were seen as a method of assessing and encouraging students to enhance their workforce skills.

Subsequently, a 1991-92 Michigan school aid act mandated that all school districts develop and maintain a portfolio for every student in grades 8-12 during the 1992-93 school year. For the 1993-94 school year, the portfolio must be implemented for all 9th

(continued on page 4)

Portfolios As Worthwhile Burdens? *(from page 3)*

graders and will be extended to 8th graders in 1994-95.

Piloted in 1990-91 by the Michigan Department of Education, the Employability Skills Portfolio is now in the first year of full implementation. More than 300 school districts have agreed to use the portfolio process to help students focus on employability skills.

Bottom-Up

Similar to the United Kingdom's "Records of Achievement," (see related article on page 6), the Vermont and Michigan portfolio programs are bottom-up approaches to assessment reform. Teachers' support is elicited through their inclusion in the development and implementation process. The states provide technical advice and some funding to assist school districts with portfolio development, but pressure as to what those portfolios must include is avoided.

Based on results in Vermont and Michigan and similar efforts in the United Kingdom, the bottom-up approach has been effective in generating ground support from teachers, principals, and school districts.

Resource Constraints

Despite Michigan and Vermont efforts to provide training to schools and districts, many time and resource constraints have been reported.

Vermont teachers felt that the greatest problem created by the portfolios "is not about what to do, but when to do it." The researchers found that over 80% of fourth-grade teachers and over 60% of eighth-grade teachers often had difficulty covering the required curriculum. And 60% of both groups reported that they often lacked the time to prepare portfolio lessons.

Teachers in Vermont also wanted more guidance. Seventy-five percent of the fourth-grade teachers and two-thirds of the eighth-grade teachers felt they lacked adequate training at least occasionally. Even teachers who had taken part in the previous pilot program reported similar needs.

"Rating diverse portfolio entries is problematic..."

In Michigan, CRESST is exploring efficient ways to describe portfolios without making an absolute judgement on portfolio contents. A portfolio *descriptive* system is under development and is based on the Michigan employability skills survey.

"Rating diverse portfolio entries is problematic," said Eva Baker and Jonathan Troper, the two CRESST researchers assisting the Michigan Department of Education. Referring to the difficulty of rating portfolios, Baker used a metaphor of student classroom performance and athletic achievement.

"Is good team performance," asked Baker, "exemplified by being the most valuable player on a team, being a player on lots of different sport teams, or by getting an effusive letter from the coach?"

Baker and Troper are helping Michigan to analyze whether *descriptive* information inside a portfolio can lead to a better understanding of student performance.

Effects on Instruction

If the portfolio resource, time burden, and rating process on teachers and schools is so great, what makes portfolios any good? Instructional effects for one thing.

CRESST/RAND researchers found that the majority of Vermont educators believed the assessment program had already had substantial positive effects on instruction. Sixty percent of the surveyed principals felt that the portfolio program had a positive effect on instruction although 25% felt that it was too early to tell. Despite this latter finding, principals seemed to agree that: "Portfolios are a worthwhile burden." According to the report:

> One relatively frequent comment (16% of principals) was that teachers increased their emphasis on problem solving and "flexible" thinking. Other principals mentioned specific changes in instructional methods or styles, including a lessened reliance on textbooks, less emphasis on drill and practice, an increased reliance on hands-on learning, increased use of interdisciplinary projects, and increased emphasis on communication of mathematics.

Teachers also supported positive instructional effects of the Vermont portfolio program: More than one-half of the surveyed teachers said they were frequently more enthusiastic about teaching math, and over 90% were more enthusiastic at least occasionally. Over 40% (of teachers) reported the following positive effects: the goals of mathematics instruction are improved; math is more closely linked to other subjects; students' attitudes towards math improve; and students are learning more math.

(continued on page 5)

Portfolios... *(from page 4)*

Another interesting instructional phenomenon was that over 80% of the surveyed teachers in the Vermont study indicated that they had changed their opinion of students' mathematical abilities based upon their students' portfolio work. In many cases, teachers noted that students did not perform as well on the portfolio tasks as on previous classroom work. This finding, supported by other performance assessment research, suggests that portfolios may give teachers another assessment tool that appears to broaden their understanding of student achievement.

Michigan teachers also reported positive effects on their students:

"We found that portfolios have a lot of educational benefits for students that aren't related to the assessment," said Catherine Smith from the Michigan Department of Education. "We're finding that one thing students begin to recognize is that their accomplishments outside of school are really important, a hobby or a club they belong to, an activity with church, or taking care of siblings. These activities have meaning for preparing them for life."

In Conclusion

Indeed, the bad news is that portfolios are definitely burdens in terms of teacher time and resources. That fact is unlikely to change. But *if* portfolios are adequately funded and lead to significant improvements in teacher motivation, instructional processes, self-evaluation, deeper understanding of content, and improved skills leading to employment, then the price may be worth it.

(See pages 10-11 for information on ordering CSE Technical Report 350, The Vermont Portfolio Assessment Program: Interim Report on Implementation and Impact, 1991-92 School Year, *Koretz, Stecher, and Deibert.* The cost is $4.00.)

NEW ASSESSMENT BOOK!

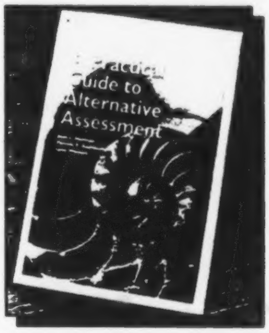
A Practical Guide to Alternative Assessment

"There is no one right way to assess students," say authors Joan Herman, Pamela Aschbacher, and Lynn Winters in their recently published book, *A Practical Guide to Alternative Assessment.* But the authors suggest that many alternative assessments offer very appealing ways to assess complex thinking and problem-solving skills. And because these new types of "tests" are grounded in realistic problems, they are potentially more motivating and reinforcing for students than traditional assessments.

Published by the Association for Supervision and Curriculum Development (ASCD), *A Practical Guide to Alternative Assessment* is written for preservice and practicing teachers, school administrators, and district and state level practitioners who are interested in creating their own alternative assessments, or in understanding the issues and improved methods for assessing student knowledge.

Within the book's 121 pages, the authors present a topical guide to alternative assessments including chapters on:

- Rethinking assessment;
- Linking assessment with instruction;
- Determining purpose;
- Selecting assessment tasks;
- Setting criteria;
- Ensuring reliable scoring;
- Using alternative assessment for decision making.



The authors discuss the development of alternative assessments within the context of a unique process model that links curriculum, learning, and instruction.

"The authors have reaffirmed the fundamental role of assessments," concludes ASCD President Stephanie Pace Marshall, "which is to provide authentic and meaningful feedback for improving student learning, instructional practice, and educational options."

Available through ASCD by calling (703) 549-9110, *A Practical Guide to Alternative Assessment* costs $10.95.

RECORDS OF ACHIEVEMENT
Lessons from the United Kingdom

> Profiling [Records of Achievement] thus arguably represents a new disciplinary technique which...has the potential to exercise more effective control than any assessment procedure yet devised.
>
> *Patricia Broadfoot*
> *University of Bristol, U.K. (1990)*

International assessment researcher Patricia Broadfoot's preceding statement is indicative of the high hopes held in many countries for portfolios, or as called in the United Kingdom, records of achievement (ROAs). In 1984, the United Kingdom, considered by many as the world leader in the development of performance assessments, mandated that records of achievement would be used in all secondary classrooms by 1990. But today, almost nine years later, despite extensive research and development, many U.K. schools are still struggling with ROA implementation issues and the U.K. government has backed away from its original 1984 requirement.

Are states and major school districts in the United States, many trying to create large-scale portfolio systems similar to the United Kingdom's ROAs, headed down a similar path?

A brief comparison between the U.K. and state and local portfolio systems indicates that, at a minimum, the U.S. already faces many of the same portfolio implementation problems as have confronted our friends from across the pond.

Commonalities

ROAs and large-scale portfolio systems in the United States are not identical. ROAs tend to have a more standardized format and cover achievement across and beyond the whole curriculum. Nevertheless, ROAs and large-scale U.S. portfolios do share many commonalities. Consider a Records of Achievement National Steering Committee (1989) that listed the purposes of the ROAs :

- to contribute to the raising of all pupils' achievement through and beyond the national curriculum;
- to improve [students'] motivation;
- to prepare [students] for the transition to further education, training and employment; and
- to help schools to consider how well their curriculum, teaching and organization enable pupils to develop their all around potential.

A comparison of these purposes to the large-scale Michigan portfolio effort (see page 3 article) indicates similar purposes. For example, the Michigan employability goals call for students to develop:

- a new and higher order of academic skills;
- Personal Management Skills that allow them [students] to develop and demonstrate the attitudes, abilities, behaviors and decision-making processes associated with responsibility and dependability;
- teamwork skills that enable them [students] to function effectively as members of multiple work teams and contribute to groups in accomplishing work tasks.

Both U.K. and Michigan goals emphasize improved student academic skills, motivation and attitude, and teamwork. Other similarities between records of achievement and many large-scale U.S. portfolio projects include the following concepts. Portfolios and ROAs:

- are valued documents meant to provide more and broader information to parents than traditional report cards;
- are owned by the pupil;
- often have as their end goal improved employment opportunities;
- feature a bottom-up design; and
- are usually implemented with minimal funding increases from mandating organizations.

If there is a major difference between ROAs and current large-scale portfolio systems here, it might be that the U.K. uses the ROAs as extended resumes for entrance into the workforce. But the Michigan Employability Portfolios could make even that difference less significant in the future, as could the demand by employers across the country to improve evidence of student workforce skills and for changing the way children are educated.

The problematic R&D issues that confront portfolio proponents in both countries are similarly overlapped. Issues of equity, issues of knowing if portfolios or

(continued on page 7)

RECORDS OF ACHIEVEMENT
Lessons from the United Kingdom *(from page 6)*

ROAs really make a difference in student learning, and issues of what makes a good portfolio have challenged both countries. Perhaps the greatest problem yet to be resolved by either the U.K. or any U.S. state is that of inadequate resources, especially teacher time. For example, the 1989 Records of Achievement National Steering Committee (1989) stated the time problem succinctly:

> It is nevertheless clear that the overall volume of teacher time is the main call [demand] that records of achievement systems make on schools' resources. In particular, tutorial time needs to be found for pupil/teacher discussions and for the preparation of summary documents. It will be essential for time to be found for in-service training for records of achievement which is carefully integrated with training needs arising from other related activities.

Both Vermont and Michigan have recognized the same major resource problems in their state-wide portfolio programs.

What's Happening Now

The 1984 United Kingdom government policy statement and the 1989 Records of Achievement National Steering Committee established a deadline of 1990 to introduce the records of achievement into all U.K. secondary schools. But according to Desmond Nuttall, one of the members of the 1989 Steering Committee and a leader of the British assessment reform movement, the government has backed away from its 1984 mandate. Overwhelmed by its recently established national curriculum and assessments, the government decided to continue its policy of voluntary records of achievement. Nevertheless, approximately 80% of the secondary schools have implemented the ROAs into their classrooms, and many primary and middle schools have followed suit.

Nuttall also reports another interesting development. The primary pressure for the ROAs has come most recently from the U.K. Department of Employment and not their Department of Education, which had previously promoted the ROAs. The Department of Employment is encouraging students to use the ROAs during job interviews. But Nuttall adds that the original plan, for all students to use the ROAs for employment, has not happened.

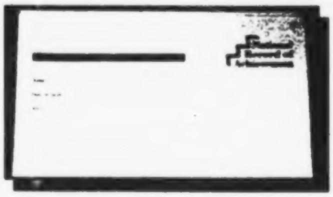
Scoring of the ROAs, or lack thereof, has not become a major issue in the U.K., as it is becoming for some large-scale assessments in the United States. Nuttall attributes this fact to their current national assessment system which does not consider ROAs to be a major test instrument.

Lessons

If there is a lesson for the United States in the U.K. experience, it might be that the implementation of portfolios is not an overnight process. Having mandated records of achievement almost nine years ago, the British are still struggling with the same types of questions faced today by state and local agencies in the United States: questions of policy, costs, training, scoring, and use by employers, teachers, and schools.

And there are many potential pitfalls that have been discussed but not fully

(continued on page 8)



What Is A Record of Achievement?

To enhance employer recognition of the Records of Achievement and to standardize the ROA basic design and contents, the United Kingdom developed a "National Record of Achievement" notebook. Each student's notebook has a nationally recognizable, standardized cover, an initial sheet describing personal details, and four main pages of records including:

- a summary of school achievements in the curriculum;
- a summary of qualifications and credits;
- a summary of other achievements and experiences; and
- a personal statement.

The Record also contains a sheet for students to record their employment history.

Compact in design and incapable of holding massive amounts of student work, the ROA focuses on evidence of student achievement including: examination results, qualifications, achievements through work experience, non-academic interests, future plans, and needs of the student. The U.K. places emphasis on students to eliminate old material.

Records of Achievement...

(from page 7)

researched. For example, if portfolios are ever to be used for such high stakes decisions (Brandt, 1992) as college entrance, they will undoubtedly have to meet the same types of stringent validity and reliability measures as current standardized assessments. Even the most ardent proponent of portfolios would have to agree that portfolios are a long way off from that level of rigor. Ultimately, the portfolio road may be a good one, but there remain many important issues to be resolved.

Our very special thanks to Professor Desmond Nuttall, Institute of Education, University of London, for providing CRESST with valuable "Records of Achievement" information.

References

Department of Education and Science and the Welsh Office. (1989). *Records of Achievement: Report of the Records of Achievement National Steering Committee.* Central Office of Information.

Brandt, R., (1992). On performance assessment: A conversation with Grant Wiggins. *Educational Leadership, 8,* 35-37.

From the Directors... *(from page 1)*

tasks rather than simply react to the tasks assigned by others. Teachers could help students accomplish significant tasks, worthy of time, reflection, and refinement. Portfolios and their evolving contents could be a source of pride for students and for their parents. As one of our relatives says, "What could be bad?"

Multiple Purposes

Although starting from the intimacy of the individual teaching/learning situation, portfolio use is rapidly transmuting to a multitude of purposes. Efficiency principles suggest we should support multiple goals met with single interventions.

But clearly not all uses will occur without side effects. For instance, should portfolios be used, as proposed in Michigan, as displays to assist employers to make a hiring decision? Why not, if they are promoted as a form of elaborated resume? Again, the process mirrors the teacher-student relationship, with relatively idiosyncratic standards applied for particular job options. The portfolio, along with other indicators, allows the employer to make a choice. The caveat is that children in different schools need to have the same level of assistance in the development of their portfolio.

Could portfolios provide an exhibit for the public and policy makers of the type of curricular emphases occurring in local schools and classrooms? Certainly a sample for review at critical grades could be made available as exhibits in displays for parents or for school board members.

But some portfolio enthusiasts have bigger aspirations. They seek to have portfolios used for student and school comparisons for accountability purposes. Portfolios would be graded to give students individual scores, to judge systems' progress toward achieving standards, and to evaluate programs. In order to accomplish these worthy ends, portfolios need to be assessed according to common standards to ensure fairness. It is the scoring of portfolios, and the concomitant stakes assigned to them, that triggers our concern.

How will portfolios be judged? If they are rated by explicit guidelines provided for their judgment, obvious consequences may occur. One is that the surface features of the scoring system will drive the portfolio development toward more superficial, and incidentally, homogeneous performance. Individual reflection and choice could be given a back seat to making sure that particular features are included, that is, buttons pushed, to get a "high" score. Students with savvy parents and teachers will surely do well.

Issues of Scoring

One alternative procedure is to use global scores, a 4 for excellent, a 3 for competent, and so on, instead of explicit scoring rubrics. Such summary scores, of course, operate against the formative, interactive strategy that portfolio assessment is supposed to promote. When global ratings are given with a lack of models or explicit criteria, what is likely to be detected are gross differences in individual talent and experience. Such scores would not help teachers to improve teaching and learning. They would function like a qualitative stanine.

Despite the frequent Olympic games allusions to judgment of qualitative performance, portfolios are in a different arena. In ice skating, for instance, even in the most creative events, there are known expectations, for example, how many

(continued on page 9)

From the Directors...

(from page 8)

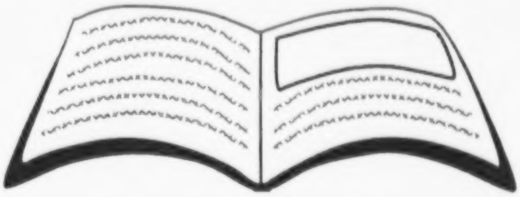
jumps of one or another type and difficulty must be made. In portfolios, the performance land of video collage, poetry, spreadsheets, and civic projects, we have no agreed upon components and few standards for any required pieces.

Do portfolios have a place in accountability? Maybe they do. On a sampling basis, portfolios can perhaps give good, qualitative information about what is happening in the best and in more typical classrooms. Do portfolios need to be scored, with all the attendant issues of reliability, valid rubrics, cost and time? Maybe; but maybe not. In a project in the state of Michigan, CRESST staff, the Michigan Department of Education, teachers, and administrators are trying to see if portfolios can be reviewed descriptively rather than scored. Our preliminary work suggests, for accountability purposes, this approach might provide sufficient information and at a reduced cost.

Appropriate Uses

Let's remember portfolios fundamentally are intended to provide qualitative information on a rich, diverse, unpredictable, and most importantly, individual set of performances. Let's not lose these key goals by converting portfolios mindlessly to inappropriate sources of quantitative information—at least not without monitoring the effects of those actions on teaching and learning. Some things are good for only one (or a few) purposes.

CSE/CRESST REPORTS



The following reports have recently been released and are available through the CSE/CRESST office. To order any report, fill out the order form on page 11, or for a complete listing of all CSE/CRESST technical reports, monographs and resource papers, please contact Kim Hurst at (310) 206-1532.

CRESST Performance Assessment Models: Assessing Content Area Explanations
Eva Baker, Pamela Aschbacher, David Niemi, and Edynn Sato, 1992, ($10.00)

Over 500 copies of this handbook have been distributed since its publication in April, 1992; consequently, we are once again promoting its use by anyone interested in developing performance assessments. Presenting a performance-based approach to assessing students' understanding of subject matter content, the handbook includes:

- a concise model of alternative assessment for those who need to develop similar assessments of their own;
- examples of successful CRESST assessment materials;
- an effective scoring rubric for performance assessments applicable to a variety of topics;
- useful benchmark papers.

The assessment model, based on a highly contextualized history performance task, requires students to engage in a sequence of assessed steps, beginning with an initial assessment of their relevant background knowledge of the particular historical period. Next students are provided with opposing viewpoints in primary source materials, typically letters or speeches of historical figures. Finally, students are asked to write an extended essay that explains the positions of the authors of the texts and to draw upon their own background knowledge for explanation.

The essay scoring rubric consists of six dimensions: a General Impression of Content Quality scale, and five analytic subscales. History experts and high school teachers have been involved throughout the study as co-designers, reviewers, and raters of the assessment and have provided valuable input into the assessment.

(continued on page 10)

More CSE/CRESST Technical Reports

(continued from page 9)

Also included in the handbook are: background information on the CRESST performance-based assessment, examples of assessments for secondary-level history and chemistry, and specifications for duplicating our technique with other topics and subject matter areas. Our rater training process, scoring techniques, and methods for reporting results are described in detail.

CRESST believes that this assessment will be useful for both large-scale applications and instructional improvement. Having used this assessment approach to research a number of technical issues in performance assessment, CRESST has evidence of the reliability, validity, and generalizability of its technique.

The Vermont Portfolio Assessment Program: Interim Report on Implementation and Impact, 1991-92 School Year
Daniel Koretz, Brian Stecher, & Edward Deibert
CSE Technical Report 350, 1992
($6.00)

See page 3 for a complete article on this report.

Design Characteristics of Science Performance Assessments
Robert Glaser, Kalyani Raghavan, & Gail Baxter
CSE Technical Report 349, 1992
($3.00)

Part of a long range goal to investigate the validity of reasoning and problem-solving assessment tasks in science, this report describes progress in analyzing several science performance assessment projects. The authors discuss developments from Connecticut's Common Core of Learning Assessment Project, the California Assessment Program, and the University of California, Santa Barbara/California Institute of Technology research project "Alternative Technologies for Assessing Science Understanding." The analysis framework articulates general aspects of problem-solving performance, including structured, integrated knowledge; effective problem representation; proceduralized knowledge; automaticity; and self-regulatory skills.

Accountability and Alternative Assessment
Joan Herman
CSE Technical Report 348, 1992
($4.00)

Despite growing dissatisfaction with traditional multiple-choice tests, national and state educational policies reflect continuing belief in the power of good assessment to encourage school improvement. The underlying logic is strong. Good assessment sets meaningful standards, and these standards provide direction for instructional efforts and models of good practice. But are these reasonable assumptions? How close are we to having the good assessments that are required?

This report summarizes the research evidence supporting current beliefs in testing, identifies critical qualities that good assessment should exemplify, and reviews the current state of the research knowledge on how to produce such measures.

The Influence of Problem Context on Mathematics Performance
Noreen Webb & Esther Yasui
CSE Technical Report 346, 1992
($4.00)

Mathematics educators and researchers argue that using realistic, complex problem-solving instruction and assessment can improve students' problem-solving skills and attitudes towards mathematics. The objectives of this study were: (a) to determine whether working with more realistic and lengthier problems during instruction will make students better able to solve similar problems on an achievement test, and (b) to determine whether the different kinds of problems (short vs. extended word problems) will provide different information about students' performance and mathematical problem-solving ability. The comparisons suggest that there are important aspects of students' ability to solve structured problems that can be measured with extended, complex, realistic problems.

Measurement of Workforce Readiness: Review of Theoretical Frameworks
Harold F. O'Neil, Jr., Keith Allred, & Eva L. Baker
CSE Technical Report 343, 1992
($4.00)

The cry of American management for workers with greater skills has spawned many commissions, task forces and studies, including five studies reviewed in this report:

- What Work Requires of Schools (Secretary's Commission on Achieving Necessary Skills);
- Workplace Basics: The Essential Skills Employers Want (American Society for Training and Development);
- Michigan Employability Skills Employer Survey;
- Basic and Expanded Basic Skills (New York State Education Department); and
- High Schools and the Changing Workplace: The Employers' View (National Academy of Sciences).

Special CRESST Report *from Robert L. Linn*

Educational Assessment: Expanded Expectations and Challenges

(1992 Thorndike Award Address)
Robert Linn
CSE Technical Report 351, 1992 ($3.50)

"Educational policymakers are keenly interested in educational assessment," says Robert L. Linn in his 1992 Thorndike Award address to the American Psychological Association. Linn points to the various attractions that assessments have for policymakers who frequently think of assessment as a "kind of impartial barometer of educational quality."

But assessments are frequently used for two questionable purposes, implies Linn, first, to point out the declining quality of American education and, secondly, as an instrument of educational reform. "Such greatly expanded, and sometimes unrealistic, policymaker expectations," he says, "together with the current press for radical changes in the nature of assessments, represent major challenges for educational measurement." Linn concludes his remarks by saying that the measurement research community must make sure that the consequences for any new high-stakes performance assessment system are better investigated than they were for previous assessment reforms.

Order Form

Attach additional sheet if more room is needed.

CSE Reports/Monographs/Resource Papers

Report Number	Title	Number of copies	Price per copy	Total Price
☐ New Videotape	PORTFOLIO ASSESSMENT AND HIGH TECHNOLOGY		$10.00	

POSTAGE & HANDLING
(Special 4th Class Book Rate)

Subtotal of	$0 to $10	add $1.50
	$10 to $20	add $2.50
	$20 to $50	add $3.50
	over $50	add 10% of Subtotal

ORDER SUBTOTAL ____________
POSTAGE & HANDLING (scale at left) ____________
California residents add 8.25% tax ____________
TOTAL ____________

Orders of less than $5.00 must be prepaid

Your name & mailing address—please print or type:
____
____
____

☐ Payment enclosed ☐ Please bill me

☐ I would like to receive a free copy of *The CRESST Line* and *Evaluation Comment* publications.

Linda Winfield Joins CRESST Staff

CRESST is delighted to have Professor Linda F. Winfield join its research staff. Dr. Winfield, formerly a principal research scientist at the Center for Research on Effective Schooling for Disadvantaged Students at Johns Hopkins University, is currently a visiting professor at UCLA's Graduate School of Education. While at Johns Hopkins, she was also co-director for "Special Strategies for Educating Disadvantaged Students," a congressionally-mandated, national study of "exemplary" urban schools.

Dr. Winfield's published work includes numerous articles focused on research and policies in urban education, including Chapter 1 evaluation, implementation and change in schoolwide projects, assessment of students from diverse populations, and equity. She has received support from the Rockefeller Foundation and the National Science Foundation for her work on literacy proficiency among black young adults.

Dr. Winfield is currently teaching the Introduction to Educational Evaluation course at UCLA's Graduate School of Education. She is collaborating with the CSE evaluation of a New American Schools Development Corporation project, "The Los Angeles Learning Centers," and will be involved in several CRESST projects involving equity and validity issues of performance assessments.

Linda Winfield

Rebuild L.A.

CRESST/UCLA contributed over $1700 to the Rebuild L.A. effort. L-R, CRESST Project Director Josie Bain, Rev. Cecil Murray and Rev. Carmen Speights from the First African Methodist Episcopal Church, and CRESST Director of Communications Ron Dietel

Center for Research on Evaluation, Standards, and Student Testing
Eva L. Baker, Co-director
Robert L. Linn, Co-director
Joan L. Herman, Associate Director
Ronald Dietel, *CRESST Line* Editor

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement number R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this publication do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

To be placed on the *CRESST Line* mailing list please write to *CRESST Line*, UCLA Graduate School of Education, 405 Hilgard Ave., Los Angeles, CA 90024-1522.

UCLA
CRESST
Graduate School of Education
405 Hilgard Ave.
Los Angeles, California 90024-1522

ADDRESS CORRECTION REQUESTED EF-16

NONPROFIT ORG.
U.S. POSTAGE
PAID
U.C.L.A.

BEST COPY AVAILABLE

Construction Versus Choice in Cognitive Measurement:
Issues in Constructed Response, Performance Testing, and Portfolio Assessment

Table of Contents and Preface

This book is based on the major presentations of a conference held at Educational Testing Service in November 1990. The first chapter explores the meanings of "constructed response" within a framework provided by validity theory. The next three chapters discuss the construct validity of constructed-response measures from psychometric, psychological, and integrated perspectives. The chapters in the following group address measurement techniques that will contribute to the incorporation of constructed-response measures into standardized assessments. Within the next group of chapters, attentions turns to discussions of more
extended assessment exercises, such as portfolios. The next chapter uses the assessment of teachers to illustrate issues in the reform of educational measurement. It provides a transition to the final chapter, which focuses on policy questions — the federal government's role, and the conflicting perspectives that influence decision making.

The book was edited by Randy Elliot Bennett and William C. Ward of Educational Testing Service.

Copies of the book are available from:

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

ISBN 0-8058-0964-3

Reproduced with permission of Lawrence Erlbaum Associates, Inc.

CONSTRUCTION VERSUS CHOICE IN COGNITIVE MEASUREMENT:

Issues in Constructed Response, Performance Testing, and Portfolio Assessment

Edited by
Randy Elliot Bennett
William C. Ward
Educational Testing Service

LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1993 Hillsdale, New Jersey Hove and London

CONTENTS

Preface ix

1 On the Meanings of Constructed Response
Randy Elliot Bennett 1

2 On the Equivalence of the Traits Assessed by Multiple-Choice and Constructed-Response Tests
Ross E. Traub 29

3 Construct Validity and Constructed-Response Tests
Richard E. Snow 45

4 Trait Equivalence as Construct Validity of Score Interpretation Across Multiple Methods of Measurement
Samuel Messick 61

5 A Framework for Studying Differences Between Multiple-Choice and Free-Response Test Items
Robert J. Mislevy 75

6 Item Construction and Psychometric Models Appropriate for Constructed Responses
Kikumi K. Tatsuoka **107**

7 Constructed Response and Differential Item Functioning: A Pragmatic Approach
Neil J. Dorans and Alicia P. Schmitt **135**

8 Item Formats for Assessment in Mathematics
James Braswell and Jane Kupin **167**

9 The Place of Portfolios in Our Changing Views of Writing Assessment
Roberta Camp **183**

10 Assessment as an Episode of Learning
Dennie Palmer Wolf **213**

11 Performance Assessment and Educational Measurement
Drew H. Gitomer **241**

12 Innovation and Reform: Examples from Teacher Assessment
Carol Anne Dwyer **265**

13 The Federal Role in Standardized Testing
Terry W. Hartle and Peter A. Battaglia **291**

14 The Politics of Multiple-Choice Versus Free-Response Assessment
Sharon P. Robinson **313**

Author Index **325**

Subject Index **331**

PREFACE

The multiple-choice question is the mainstay of standardized testing programs in the United States. The format has achieved this position because it permits inexpensive and apparently objective scoring; because such questions can be answered quickly, allowing broad content coverage within a testing session; and because a sophisticated statistical technology has evolved to support the analysis and interpretation of test results.

The reliance on multiple-choice questions, however, is increasingly criticized. Many have argued that tests and, in particular, test formats significantly influence education. Multiple-choice assessments are said to encourage the teaching and learning of isolated facts and rote procedures at the expense of conceptual understanding and the development of problem-solving skills. It is believed that, for education reform to occur, the methods used to measure attainment must themselves be transformed.

To address the limitations of the multiple-choice format, many educators and psychologists have advocated increased use of constructed-response tasks. These tasks may be as simple as producing a numerical answer to an arithmetic question or as extensive as producing the numerous drafts that culminate in a finely honed essay or planning and conducting a series of scientific experiments. Proponents argue that constructed-response assessments, especially those that require extended problem solving and yield complex productions, measure different skills and promote deeper learning than do multiple-choice measures.

The use of such tasks, however, raises several critical concerns. If the answer to a question is an extended problem solution, fewer questions can be asked in a fixed testing period, reducing the breadth of content coverage possible. The less constrained the task and the solution, the greater is the possibility that lack of standardization in test administration, and lack of objective criteria for evaluation, may adversely affect the comparability of results across persons and situations. These conditions can threaten the representativeness of the test results as a sample of the individual's capabilities, and thus the validity and fairness of the test.

Growing attention to these and related issues has suggested that it would be timely to bring together persons who could contribute to an understanding of problems and possibilities associated with the various assessment formats. First a conference, and then this volume, resulted.

The conference, sponsored by Educational Testing Service, was held in Princeton, New Jersey, in November, 1990. Speakers and attendees represented a variety of viewpoints in educational research and policymaking. The presentations and discussions were informative, provocative, and notably lacking in polemics.

This book, comprising nine chapters based on the major conference presentations plus five newly invited contributions, maintains the same tone. Rhetoric calling for the abolition of traditional testing methods as useless or pernicious, or on the other hand for dismissal of new approaches as impractical, is lacking. Rather, the authors seek to provide perspectives and build frameworks that will contribute to future research agendas and policy debates. Such statements are not as dramatic as the more extreme positions that can be found in the press and even in journals, but they are, we believe, more useful.

The first chapter in the volume, that by Bennett, explores the meanings of "constructed response" within a framework provided by validity theory. The next three chapters discuss the construct validity of constructed-response measures. Traub provides a psychometric perspective; Snow, a psychological one; and Messick, an integration of the two.

The chapters in the following group address measurement techniques that will contribute to the incorporation of constructed-response measures into standardized assessments. Mislevy outlines the use of "inference networks" in evaluating the contributions of different types of test questions. Tatsuoka discusses a model for item design to elucidate the skills and knowledge underlying observable performance. Dorans and Schmitt describe techniques for the analysis of group differences in item performance. Finally, Braswell and Kupin examine alternative formats for assessment in mathematics.

With the next group of chapters, attention turns to discussions of more extended assessment exercises. Camp explores the role of portfolios in the assessment of writing. Wolf draws from both the classroom and the reflections of practicing artists to view assessments as occasions of learning. Gitomer provides a framework for the design of performance assessments in educational measurement.

Dwyer's chapter uses the assessment of teachers to illustrate issues in the reform of educational measurement. It provides a transition to the final chapters in the volume, which focus on questions of policy—Hartle and Battaglia from the perspective of the federal government's role, and Robinson exploring the conflicting perspectives that influence decision making.

Important contrasts between the more narrowly psychometric and the social policy perspectives are evident in these chapters. The two viewpoints are in agreement in seeking means of improving educational measurement; but they differ, at least implicitly, in what is meant by "better." From the policy perspective, better measurement involves tasks that have verisimilitude, that send the right messages to those concerned with education, and that help directly and indirectly to cause increased success for learners. From the psychometric, "better" means more reliable or more representative of cognitive skills underlying an achievement, or perhaps less susceptible to contamination by construct-irrelevant group differences. From the first of these perspectives, it may make good sense to trade some accuracy of measurement for a superior assessment; from the second, that proposition is almost a contradiction in terms.

Another aspect of the contrast in perspectives is that there are significant differences in how the line is drawn to distinguish variations in measurement methodology that make a difference. From the psychometric viewpoint, the step from a multiple-choice mathematics question to one in which the examinee is asked to grid an answer in is a very big change; one has to be concerned about the consequences of this change for test reliability, difficulty, speededness, and so on. Any variation in format and scoring rubric must be studied exhaustively. From the policy perspective, however, such changes are minor. The constructed-response measures that are seen as likely to make a difference are far more complex and real-worldly, barely on the same continuum with the array of measures likely to be considered by those for whom such factors are the critical concerns.

Just as evident as the differences should be the indication of ways in which these contrasts might be bridged. Several of the chapters offer organizing schemes and discussions that can begin the synthesis needed to promote the objective shared by all of the contributors: achieving more socially useful, socially responsible measurement. We hope this volume contributes, if only in a small way, to that important goal.

Randy Elliot Bennett
William C. Ward

POLICY INFORMATION CENTER

EDUCATIONAL TESTING SERVICE • PRINCETON, NJ 08541 • 609-734-5694 • MAIL STOP 04-R

BEST COPY AVAILABLE